DOCUMENT RESUME

ED 340 771 TM 018 027

AUTHOR	Bunderson, C. V.; And Others
TITLE	Computers in Educational Assessment: An Opportunity To Restructure Educational Practice.
INSTITUTION	Institute for Computer Uses in Education, Pennington, NJ.
SPONS AGENCY	Congress of the U.S., Washington, D.C. Office of Technology Assessment.
PUB DATE	21 Dec 90
NOTE	81p.; Contractor report prepared for the Office of Technology Assessment titled "Testing in American Schools: Asking the Right Questions." For related document, see TM 018 025.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	Computer Assisted Instruction; *Computer Assisted Testing; Computer Software; *Educational Assessment; Educational Change; Educational Technology; Elementary Secondary Education; Formative Evaluation; *Management Information Systems; Models; Recordkeeping; Test Construction

ABSTRACT

This paper examines ways in which hardware and software technologies can be used for effective educational assessment. The analysis considers current uses of computer technology in educational assessment and future applications. Computer systems can integrate administration of measurement instruments, presentation of instructional materials, recordkeeping, and management of instructional activities. Computers can provide a new kind of growth environment for classrooms and learning laboratories, and their introduction is a promising way to stimulate productive learning. Models that do not involve teachers, other educators, and students in slow growth will probably not work. It is best to start where users now are, and introduce formative evaluation as a fundamental aspect of implementation. Eight specific recommendations are given for increasing computer use in schools. Three recommendations dealing with the purposes and methods of testing in the schools are: (1) greatly increase the frequency and variety of help services compared to high-stakes assessments, but balance the two; (2) greatly increase the frequency of formative evaluation, and provide funding and incentives to use the evaluation data for ongoing improvement of educational programs; and (3) increase the use of alternate methods of assessment (i.e., that require human judgment and that measure more complex, integrated, and strategic objectives.) Three recommendations dealing with the new infrastructure for Computerized Educational Assessment (CEA) are: (4) foster new item types and uses of portable answer media in order to utilize the current testing infrastructure more creatively; (5) encourage the development of a localized infrastructure of Integrated Learning and Assessment Systems, and the coordinated evolution of central sites for development of help systems and tests, and for research and development; and (6) encourage the professional development of teachers and other professionals who are knowledgeable and skilled about both the human judgment and the technical aspects of CEA, and are skilled at integrating assessment with instruction. Recommendations dealing with policy are: (7) federal and state policy should both provide research and development funds and stimulate private sector investment in improving technology-based assessment practices; and (8) high professional testing standards must be maintained and must continue to evolve for CEA systems. Eight tables present study data, and two illustrative figures are included. (SLD)

COMPUTERS IN EDUCATIONAL ASSESSMENT

An Opportunity to Restructure Educational Practice

A paper prepared for

U.S. Congress
Office of Technology Assessment

by

The Institute for Computer Uses in Education
(ICUE)
106 West Franklin Avenue
Pennington, NJ 08534

C.V.Bunderson
J.B.Olsen
A.Greenberg

December 21, 1990

This contractor document was prepared for the OTA assessment entitled *Testing in American Schools: Asking the Right Questions*. This is being made available because it contains much useful information beyond that used in the OTA report. However, it is not endorsed by OTA, nor has it been reviewed by the Technology Assessment Board. References to this should cite the contractor, not OTA, as the author.

EXECUTIVE SUMMARY

This paper was commissioned by the Office of Technology Assessment (OTA) of the U.S. Congress in August 1990 as a part of a major review of educational assessment. Of several commissioned projects, this one dealt with computer technologies and their application in educational assessment.

One problem motivating the OTA statement of work is that computer advances have not as yet led to fundamentally new paradigms and approaches to measurement of human cognitive functioning, ability, or achievement. The key research question addressed by this paper is how can hardware and software technologies be marshalled for effective educational assessment? Can such computer applications reduce some of the testing problems inherent in current methods of educational assessment? Current testing methods have been criticized in many ways, including the failure to assess adequately thinking and problem-solving processes. These processes are vital to a citizenry who must compete in an increasingly complex technological environment, and in an intensely competitive world. Furthermore, as a senator and congressman who encouraged this study asked, do not current tests (which frequently emphasize minimum competencies rather than complex thinking) exert an undue influence on what we teach and how we teach in our schools?

The Statement of Work required that this general problem area be approached through a two-fold analysis: (1) Current Uses: How is computer technology currently being used for the assessment of the various objectives of assessment? (Consider the benefits and limitations of each of the technologies currently being used.) (2) Future Uses: Looking ahead, what emerging software and hardware technologies may have implications for educational assessment, both through extending existing methods of assessment and through generating completely new ones?

The two most extensive of the four sections of this paper, sections II and III, meet these requirements of the statement of work for analysis of current practices and future possibilities. In addition, Section I sets the stage and Section IV presents a summary, conclusions and recommendations.

FINDINGS: THE ACTUAL AND THE POTENTIAL

Educational measurements may be administered using either portable answer media or interactive testing stations. Portable answer media include answer sheets, readable on optical scanning equipment, or portable keypads or barcode readers which facilitate the entry of a letter or a number. Current practice is dominated by the use of printed tests with scannable answer sheets.

Computerized Administration: When computers are used to administer a test, we lose portability and must install computerized workstations or special simulation devices in a learning center or assessment center. Portable computers, including notebook computers, may be used in the future for administration of educational measures at temporary locations.

Computers may be used in any of several processes during the life cycle of an educational measurement instrument. They may be used to aid in:

1. design and development of measurement instruments;
2. distribution of measurement instruments to testing locations;
3. administration of the measurement instruments; and,
4. analysis and record-keeping after administration.

Computers can improve and even transform any of these processes, but the emphasis in this paper is on the processes of administration. When the measurements are administered by computer, it is likely that computers are used extensively in the development, distribution, and later analyses as well.

A model for technology diffusion used by OTA distinguishes three levels of penetration of a new technology.

1) Substitutive
2) Incremental
3) Transformational

These three levels provide a useful framework to report the findings. Computer administration may be used as a substitute for conventional test administration, which uses scannable answer sheets, printed test booklets, and occasionally, adjunct audio and visual media or objects which the test-taker manipulates.

Testing has not always been dominated by printed, group-administered formats made up of a goodly number of short items. Albert Binet's pioneering intelligence test developed early in this century was individually administered. It provided the test-taker with a variety of standardized tasks that could be answered vocally, or through a physical performance. Many useful clinical tests today are individually administered, human-judged tasks. The Army Alpha, an intelligence test used to screen recruits, represents the first widespread use of group administered paper and pencil test consisting of multiple choice items. Such tests are wholly objective – that is, they do not require assessment in the judging of responses to individual items. A distinction is made between assessment and measurement in this paper. Assessment requires human judgment to interpret observed responses and assign scores, and to make decisions based on the scores and other information. By contrast with this costly procedure, group-administered paper and pencil tests provided simple scoring rules. The number of correct items is typically summed and a simple scoring formula is applied to correct for guessing.

Substitutive Uses of Computers

It is perhaps not surprising that the current uses of computers in educational measurement are primarily substitutive for paper and pencil item tests, rather than individually administered tests, since the paper and pencil tests represent the dominant format for educational measurement. Substitutive applications of computers take away the answer sheet and test booklet and present the items on an electronic computer display, receiving the responses from a keyboard or pointing device. These substitutive tests use the same kind of scoring rules used in the paper versions.

The use of such tests brings many benefits, including greater efficiency and standardization, and a closer link of achievement tests to instructional modules, both in time and in content. Even when separated from the instruction as separate pre- and post-tests made up of verbal items, the integration with instruction is much tighter. When delivered by the same interactive computer system used to provide instruction, hints and helps following testing can be given at the moment of interest and of need.

Weaknesses of Conventional Tests. Unfortunately, the wholly substitutive use of computers cannot transcend the measurement limitations of the original paper and pencil tests, nor will it eliminate the negative consequences if computerized tests are still administered as separate and aversive activities not integrated with instruction. Item tests have been criticized for measuring shallow factual verbal knowledge and being less effective at measuring integration, organization, synthesis, problem solving, design, creativity, motivation, persistence, etc. More damaging, when such tests are imposed as measures of minimum competencies in reading and math and schools are held accountable for bringing up scores on item tests, they become the focus of instruction. Boring drills to prepare for the tests may monopolize the time that would be given to inherently more interesting subjects and to thinking, problem solving, producing, etc.

Incremental Improvements Over Substitutive Computer Tests

Several incremental improvements over computerized conventional tests are currently in use. The two most widely used systems are called computer adaptive and computerized mastery tests. Unlike the computerized conventional tests that give a fixed number of items in a fixed order (like the paper tests they substitute for) the incremental improvements change the sequence and the number of items dynamically. These tests also use conventional items, but in an adaptive test, each new item is selected depending upon how the student is doing, with more able students getting more difficult items and less able students getting easier items. Computerized mastery tests administer a series of item clusters, stopping when a decision has been reached about whether the test taker is above or below some pre-established cut score. Both take far less time for more accurate measurement.

Other incremental improvements aim to derive more information of a diagnostic nature out of the student's responses to the items. These systems combine cognitive analysis with measurement in a promising way, but are highly experimental.

Improved display and response processing offer many incremental advances. Computer graphics, color, animation, video and audio add interest and realism to the item types possible in computerized measurement.

Transformational Uses

As the OTA statement of work suggests, there are transformational possibilities inherent in the new interactive computer technologies and multi-media display capabilities. Not only can items currently transcend the limited verbal representation on the printed page through multi-sensory presentation, but the definition of an item as a short, quickly administered task can be replaced by standardized performance tasks of longer duration which require significantly more integrated processing. These complex and integrative tasks can be presented as simulations or games, or as tool-like laboratory environments that require exploratory, constructive, and hypothesis testing activities on the part of students. Such highly integrated and conceptually demanding tasks, when used in assessment, can truly be labelled as transformational. Unfortunately, such tasks are not yet having much impact in educational assessment. They are slowly coming into use in experimental instructional settings. They can be scored holistically by human judges, but the psychometrics for scoring computer-collected responses to them and determining their reliability and validity of such scores has not been developed. The infrastructure for developing, distributing, and interpreting them is not in place.

In another potentially transformational use of computer technology in assessment, students are provided with computer tools for designing, developing and producing "exhibits": written documents, presentation materials, performance scripts or plans, or other products of mind. Software productivity tools exist for multi-media design, editing and production of student exhibits. These can be used in combination with project assignments and with portfolio assessment methodologies to get at learning objectives that deal with creative production. In portfolio assessment methods, students use portfolios to manage their own set of exhibits, including the intermediate stages in the development of some of the exhibits. Such uses of technology can be truly transformational in both instruction and assessment because they provide a penetrating method to assess *process* as well as product. Using holistic assessment based on human judgment, the students and teachers can discuss intermediate products and can consider the issues of strategy and tactics leading to the production of a polished final product. An exciting prospect for the future is that student responses can be recorded while they are using computer design and production tools. Future intelligent software can be used to provide hints and helps to improve strategy and tactics at the moment of need.

Another transformational possibility is that the use of productivity software tools may be augmented through access to large data bases of images, audio and text. Software for searching such data bases can be used in the production of student exhibits. These potential transformational uses of computers are gaining momentum in the schools as a part of instruction in writing and other subjects, but have not yet been formalized as a new assessment method.

Transformation of Whole Systems. Potentially the most powerful of transformational possibilities inherent in the idea of computerized educational assessment lies in transforming the relationship of assessment to instruction and to instructional management. Educational measurement has been viewed as an activity separate from instruction. It precedes an educational sequence or follows it, but is left up to the teachers to prepare and use their own measures during an educational sequence. The patterns teachers may look to in making up their own measurement instruments emphasize judging and grading rather than helping and guiding. In its most visible embodiments, centrally prepared other than teacher-prepared, Educational Measurement is viewed with fear and dislike, and is not seen as integral to the processes of learning and teaching.

Printed materials that integrate assessment with instruction and can aid in instructional management are possible, and some have been developed, but computer-based systems that integrate instruction, educational measurement, and management offer a major transformational step beyond these. Such systems can transform the roles of educators and their effectiveness. Their roles as decision-makers can be enhanced by supplying them with continuous and up-to-date information to guide their interpretations and decisions. Students can become active problem-solvers, strategists, and producers rather than passive recorders and regurgitators.

Help Systems Versus High-Stakes Tests. A computerized assessment system that is fully integrated with instruction and is never used for grading or high-stakes accountability is called a "help system." It is distinguished at some length from a high-stakes test in Section I. If it is true that the nation needs to use more school time for progress and growth, and less for grading, judging and measuring minimums, then help systems which integrate assessment and instruction (and are never used for grading either the students or the teachers) offer a promising alternative.

Despite slow progress toward transformational and even incremental forms of computerized assessment, the message of this report is positive and hopeful -- that Computerized Educational

Assessment (CEA) offers great potential for the assessment of individual learners, for the evaluation of educational programs, and for the transformation of professional roles and practices of educators.

Educational Measurement Infrastructure

The key research question stated in the statement of work is: "How can increasingly powerful hardware and more sophisticated software be marshalled for effective educational assessment?" To answer this question, the nature of the educational measurement infrastructure must be understood. Two aspects of this infrastructure are discussed in the paper: the technological base and the human talent. The infrastructure of the current system is centralized. Large computer facilities with paper processing equipment are at the center. These facilities are managed by testing companies, state education agencies, and in some cases, professional organizations who test for licensing and certification. There is a human infrastructure of part-time test administrators for large national testing organizations, usually made up of college teachers and secondary school teachers. They use borrowed or leased facilities for a few hours to administer the tests under controlled conditions. Unfortunately, there is no permanent physical location for computerized assessment (even assessment intimately integrated with instruction) and no trained talent base to administer the assessments at the point when it could benefit the learning and thinking processes of school children. By way of contrast, colleges and many high schools have admissions offices and guidance offices which interpret educational measurements to make admissions decisions and placement decisions, and to provide guidance and counselling.

Since the current infrastructure for educational measurement is already in place, the question arises as to whether it can be revitalized with a host of new item types and new methods that use printed and scanned answer sheets more creatively. This paper reports on exciting new printed item types that can revivify and improve the current testing infrastructure. Students can mark out words in running text, identify objects or structures by marking on printed pictures, can draw arrows, lines or simple graphs, and can even print numbers and letters that can be recognized by the optical scanners and associated software.

A new infrastructure is needed in the schools, consisting of networked computer workstations, in order to obtain the transformational benefits of computer uses in assessment. This equipment is necessary to introduce computer-administered standardized performance tasks, including simulations and games, and process measures during tool use in the production of student exhibits. In addition to capitalizing and installing the technology base, a new human talent base must be developed. Such a new local infrastructure is being installed in the schools even now. There are two kinds of systems being installed: Integrated Learning Systems (ILS) and networked computer labs for tool use. The dominant mode now evolving uses a relatively powerful file server computer with large central disk file capacity for curriculum and for record-keeping. These systems may be used for computer-aided instruction; they may also be used as labs for productivity tool use. The integrated learning systems come closest to the environment foreseen in this paper as providing an infrastructure for the delivery of computerized educational assessments integrated with instruction. The recommended system is referred to in this paper as an ***Integrated Learning and Assessment System (ILAS)***. This is a concept, not a particular product. It is an evolutionary improvement over the Integrated Learning Systems (ILS) and the computer labs that now constitute an important growth industry in education.

By integrating assessment, instruction, and records management, the ILAS offers the most profound transformational opportunity of the computer. It provides an environment where sensitive and important assessment, in the full sense of the term defined herein, can take place and in association with good management and instructional decision-making. These decisions are also enhanced by the properties and features of the ILAS. Skilled teacher/managers, who are also good assessors and decision-makers, can develop professionally over time in a context of teaching with the use of an ILAS. This provides a challenging and promising professional growth path for teachers toward new roles more consonant with national needs. In achieving this professional growth, teachers working with Integrated Learning and Assessment Systems will be greatly strengthened in their efforts lead students to high standards of achievement in the complex and demanding objectives needed for success in a technologically intensive and internationally competitive world.

RECOMMENDATIONS

The eight recommendations developed as a result of the analysis of current trends and future possibilities provide a vision and a direction. Specific details will need to be worked out state by state, district by district, school by school, and company by company within the rapidly growing ILS industry. This industry, which has an annual volume of over $800 million, is now evolving and will continue to evolve toward the integration of measurement and assessment with instruction (ILAS). Because of the generality of the recommendations, they have implications for research, development, and implementation support. There are substantial policy implications for these recommendations as well, for they lead toward transformational innovations. In particular, the entire Integrated Learning and Assessment System (more properly, Integrated Learning, Assessment and Formative Evaluation System) that is envisioned in this paper is transformational of the roles and activities of educators, both administrators and teachers, and of students. There are power shifts as students and teachers are given more control over the information they need to make decisions, and more control over the resources and policies that are connected with the decisions they will need to make to optimize and manage the learning enterprise. Principals and other administrators must become instructional leaders; interpreters of decision-oriented data gathered from the high-stakes assessments and formative evaluation data.

The recommendations are divided into two major groups, and there are three recommendations in each group. The first set deals with the purposes and methods of testing in the schools, and the second set deals with building the needed infrastructure for computerized administration of improved educational measurements.

Three Recommendations Dealing With the Purposes and Methods of Testing in the Schools

Recommendation One: Greatly increase the frequency and variety of help services compared to high-stakes assessments, but balance the two.

Recommendation Two: Greatly increase the frequency of formative evaluation, and provide funding and incentives to use the evaluation data for ongoing improvement of educational programs.

Recommendation Three: Increase the use of alternate methods of assessment (i.e., that require human judgment and that measure more complex, integrated, and strategic objectives.)

This recommendation is explained in terms of a conceptual framework consisting of four measurement methods and four kinds of objectives. The following table summarizes Recommendation Three, which calls for research, development, and implementation that emphasizes the measurement methods of performance tasks, exhibits, and process measures during tool use, and thus more teaching and learning of the higher-order constructs of integration, creative production and strategies. The use of item tests for scaffolding knowledge should be improved and extended beyond scaffolding knowledge (verbal knowledge about some topic -- terminology, definitions, classifications, simple role use) as far as possible.

Recommendation 3: Measurement Methods Suitable for Different Objectives

TYPE OF ACHIEVEMENT OBJECTIVE	MEASUREMENT METHOD			
	Item Test	Standardized Perf. Tasks	Student Exhibits	Process Measures
Scaffolding Knowledge	●●●●	●●	●	●
Integrated Performance Capability	●	●●●●	●●	●●●
Creative Production Capability	●	●●	●●●	●●●
Strategy Improvement	●(-)	●●●	●●	●●●●

The greater the number of asterisks in a cell, the more appropriate a particular measurement method is for the class of objective listed to the left. Thus, item tests are the best suited for measuring scaffolding objectives. In general, performance tasks offer the greatest promise for measuring integration objectives and process measures the greatest potential for measuring strategies.

Three Recommendations Dealing With the New Infrastructure for Computerized Educational Assessment (CEA)

Recommendation Four: Foster new item types and uses of portable answer media in order to utilize the current testing infrastructure more creatively.

Portable answer media (mainly answer sheets) should be freed from domination by the multiple choice item type. It is now possible to develop and introduce many new item types and task types for paper delivery in help systems as practice and feedback worksheets, and to improve testing of high-stakes outcomes. In short: use the infrastructure that is in place to broaden the assessment options available.

Recommendations Five and Six deal with building the new technological and human infrastructure for computer-administered assessment.

Recommendation Five: Encourage the development of a localized infrastructure of Integrated Learning and Assessment Systems, and the coordinated evolution of central sites for development of help systems and tests, and for R&D.

Recommendation Six: Encourage the professional development of teachers and other professionals who are knowledgeable and skilled about both the human judgment and the technical aspects of CEA, and are skilled at integrating assessment with instruction.

Recommendations Dealing with Policy

Recommendation Seven: Federal and state policy should both provide R&D funds and stimulate private sector investment in improving technology-based assessment practices.

Recommendation Eight: High professional testing standards must be maintained and must continue to evolve for CEA systems.

TABLE OF CONTENTS

EXECUTIVE SUMMARY i

FINDINGS: THE ACTUAL AND THE POTENTIAL

RECOMMENDATIONS iv

Three Recommendations Dealing With the Purposes and Methods of Testing in the Schools v

Recommendations Dealing With the New Infrastructure for Computerized Educational Assessment (CEA) vi

Recommendations Dealing with Policy vi

SECTION I: FRAMING THE ISSUES 1

BACKGROUND: ISSUES IN MEASUREMENT AND ASSESSMENT 1

Contrasting Views Toward Measurement 1

A Distinction between Measurement and Assessment 2

Evaluating America's Educational Needs 2

Has an Historical Shift Occurred 3

PURPOSES AND OBJECTIVES OF EDUCATIONAL MEASUREMENT AND ASSESSMENT. 3

Contrasting High-Stakes Assessment and Help Services 5

Purposes and Objectives of Program Evaluation 8

METHODS FOR MEASURING THE ACHIEVEMENT OF INDIVIDUALS 10

Item Tests 10

Standardized Performance Tasks 10

Creative Products (Exhibits) 11

Indirect Measures During Tool Use 11

SUMMARY 11

SECTION II: CURRENT USES OF COMPUTERS IN ASSESSMENT 13

TYPES OF COMPUTERIZED MEASUREMENT SYSTEMS 13

Portable, Non-Interactive Answer Media 13

Interactive Testing Stations 13

Essential Processes in Administering an Educational Measurement Instrument 13

ADVANTAGES AND DIFFICULTIES OF ANSWER SHEET SYSTEMS 15

Minimum Competency Testing and Its Effects 16

SUBSTITUTIVE, INCREMENTAL, AND TRANSFORMATIONAL APPLICATIONS OF INTERACTIVE TESTING STATIONS 17

TABLE OF CONTENTS, second page

SECTION II (cont.)

INCREMENTAL IMPROVEMENTS IN COMPUTER-ADMINISTERED TESTS 20
- Computerized Adaptive Testing 20
- Computerized Mastery Tests 22
- Computer-Based Diagnostic Testing 23
- Incremental Improvements to the Display of Item Contents 23

POTENTIALLY TRANSFORMATIONAL APPLICATIONS OF COMPUTERS 24
- Current Uses of Computers to Administer Standardized Performance Tasks 24
- Current Uses of Student Products in Assessment 26
- Process Measures During Tool Use 26
- Computer Applications That Integrate Assessment With Instruction 26

USE OF COMPUTERS IN PROGRAM EVALUATION 28

CURRENT USES OF COMPUTERS IN DEVELOPMENT, DISTRIBUTION, AND ANALYSIS OF EDUCATIONAL ASSESSMENTS 30

THE INFRASTRUCTURE FOR EDUCATIONAL MEASUREMENT AND ITS PROBABLE EVOLUTION 31
- A Critique of the Educational Measurement Infrastructure and the Current Testing Industry 32

SECTION III: FUTURE USES OF COMPUTERS IN ASSESSMENT 39

TWO FUTURE SCENARIOS FOR COMPUTERIZED EDUCATIONAL ASSESSMENT: 39

SCIENTIFIC TRENDS 40

FUTURE PROGRESS IN INSTRUCTIONAL SCIENCE AND MEASUREMENT SCIENCE DEPENDS ON THE DEVELOPMENT OF GENERALIZED ACHIEVEMENT CONSTRUCTS 43

TECHNOLOGICAL TRENDS 47

ADMINISTRATION OF FUTURE CEA SYSTEMS. 48

CEA SYSTEMS WILL EXERT A TRANSFORMATIONAL INFLUENCE ON ALL PERSONNEL IN THE EDUCATIONAL ENTERPRISE 58
- How CEA Systems Will Impact the Role of Teacher 58
- Barriers to Widespread Implementation of CEA System 59

SUMMARY 60

SECTION IV: FINDINGS AND RECOMMENDATIONS 61

FINDINGS: CURRENT USES OF COMPUTERS IN ASSESSMENT. 61

RECOMMENDATIONS 64
- Recommendations Dealing With the Purposes and Methods for Testing in the Schools 64
- Recommendations Dealing With the New Infrastructure for Computerized Educational Assessment (CEA) 65
- Policy Recommendations: 66

IN CONCLUSION 67

SECTION I: FRAMING THE ISSUES

BACKGROUND: ISSUES IN MEASUREMENT AND ASSESSMENT

How is measurement viewed and used in Education vs. other occupations? What are the differences between measurement and assessment, and the shifting social roles of each?

Contrasting Views Toward Measurement

We scarcely notice how measurement standards for time, quantity, distance, velocity, and money structure our daily communications and commerce, guiding our decisions. For scientists and engineers, measurement is an inseparable ally, used to calibrate and test equipment, quantify observations, validate hypotheses, theories and designs. For skilled operators of complex systems, measurement provides an appropriate array of instrument readings indispensable to timely decision making. Business decision makers measure the performance of their business unit, summarizing it in terms of revenue, expense, and productivity. The quality of their strategic decisions depends on the accuracy and completeness of their data. For professional artists and athletes it is the feedback of coaches during practice, and scorecard statistics that are indispensable to improving performance. In these occupations, however, measurement is as familiar and unobtrusive as are the common standards for the lay public. It is a totally integrated aspect of work, eagerly sought, always essential to sound decision making.

In contrast, educational measurement is less integrated and more intrusive. It has become a coercive tool of the educational system to assure compliance in the completion of assignments and studies, an embarrassing index of intelligence or illiteracy, a feared gatekeeper of opportunity, a weapon used by administrators to indicate accountability and academic productivity. Teachers and administrators may resist the introduction of measurement, but in this they are not so different from people in other occupations. Few employees or professionals are enthusiastic about measurement when they are being evaluated in bureaucratic settings by persons external to their work group. When measurement is intrusive, high stakes and threatening, it is not welcome.

Is there a kind of measurement system in education that would evoke less resistance, dislike and fear than the current system? Could measurement become a more integrated and helpful aspect of the work of administrators, teachers and students? Certainly there must be a way, and it will have to involve a move toward more helping and less judging.

America needs a highly professional educational workforce. One indisputable mark of professionalism in many occupations is the ability to interpret measures essential to their work with sensitivity, balance, and good judgment. Timely and appropriate information is as important for educators as for others to support and clarify their decisions. The feedback provided by helpful measurement is not only vital for immediate decision making, but also for improving future decisions based on a framework that enables one experience to be compared with another. Without clear indicators of desired educational objectives, educators orient their priorities toward unspoken and unwritten priorities, including those unrelated or even inimicable toward student learning.

New developments and technologies are vitally needed in Educational Assessment. The invisible intellectual acquirements of students are very hard to measure, and it is even harder to do so in a timely and up-to-the-minute way. The performance and productivity of an educational system, whether single classroom, a local school, a district, or a state, is also hard to measure; especially using the paper and pencil instruments which now predominate in measurement practice. Educational measurement experts devise scales to make visible many subtle nuances inherent in learning, problem solving, and thinking. In the hard sciences, the more deeply embedded and invisible is an hypothesized construct, the more costly it is to measure, requiring complex instrumentation. Might it be possible to equip the classrooms of this nation with computerized assessment instrumentation that will enable measurement to become the inseparable and unobtrusive ally of education? The answer this paper gives is "yes, but...". Yes, there are great opportunities inherent in current and future uses of computerized measurement instruments. But significant R&D is needed, a new infrastructure must be put in place, and thoughtful policy development must guide both.

A Distinction between Measurement and Assessment[1]

Taking measurements is not the same as performing an assessment. An assessment requires interpretations and human judgments, while a measurement alone does not. Consider an example from medical practice: A pediatrician may measure very precisely the height and weight of a child, but to interpret those measurements and take action requires an act of judgement: an assessment. There is no unambiguous diagnosis of overweight or underweight based on the measurements alone. Other factors such as age, sex, genetic factors, and health history may be even more important than the measures for a given diagnosis and prescription. Especially valued is subtle clinical judgment based on years of experience in interpreting the appearance, demeanor, odor and other subtle cues.

A particular assessment can be valid or invalid. Consider the sequence of activities that follow the measurement of some set of human attributes.

measures -> interpretations -> decisions -> consequences

The act of measurement is neither valid nor invalid (it may be appropriate or not, or accurate or not), but the interpretation that follows the measurement may be valid or invalid. These judgments should employ more information than is available from a single measurement, especially when the decision involves high stakes to some person or persons. Similarly, the decision is a judgment based on the interpretation and other information, and on the options available. A decision has consequences, both negative and positive, which must be considered by the decision maker. Good tests provide statistical evidence that the measure will predict future outcomes. Knowing this, the decision maker can be more confident that certain positive outcomes are more likely than negative ones. Good tests also provide published evidence that a decision will at least be as fair and equitable as possible when negative consequences for some group of test-takers follow.

Assessors and evaluators cannot escape the responsibility for making valid assessments based on all the evidence available. For assessment to be used in a more powerful manner in schools, far more training and sophistication will be needed, especially when the assessments are high-stakes. Evaluators should not be permitted (or coerced by external threats) to shrug their shoulders and "let the measurement make the decision."

Evaluating America's Educational Needs

America's policy makers have read a variety of measures and indicators and know that we are "a nation at risk."[2]. Consider a hypothetical national scale of educational achievement. At the bottom of the scale is a large number of educational drop-outs and failures. A disproportionate number are poor, or from ethnic and cultural minority. The middle range of our hypothetical scale also presents a grim picture. Those who successfully enter the workforce may not be sufficiently skilled to enable our economy to participate effectively in an increasingly competitive global market. At the top of the educational achievement scale the situation is positive - top colleges and universities in the United States are "world class". The fact that increasing numbers of faculty and graduates are foreign-born may be taken as an indication that America's candidates for higher education are competitive, but it is also an indication that our universities are sought out by the best minds in all nations.

[1] The terms *Educational Measurement* and *Educational Assessment* are defined more formally in other publications. In educational publications, the term assessment has a variety of uses. The emphasis in this paper is on making valid interpretations and decisions for action from measurement. The concept of validity involves both the interpretations and the consequences of action, not just the quality of the measure itself. [See S. J. Messick, "Validity", in R. L. Linn (Ed.), Educational Measurement, Third ed., (New York: Macmillan, 1989), pp. 13-103.]

[2] D. P. Gardner, A nation at risk: The imperative for educational reform. (Washington, DC: National Commission on Excellence in Education, 1983).

Moving from the hypothetical to the actual, there is recurring and growing criticism about the standardized educational tests we use to measure achievement. These criticisms include charges that the tests do not address higher order thinking skills, adaptability to new circumstances, or creativity. Instead, they are seen to measure disconnected snippets of knowledge that are soon forgotten and are not integrated into a knowledge structure that can be used, along with powerful thinking and reasoning skills, to deal with the complex technology and new learning demanded of world class workers and citizens[3].

Has an Historical Shift Occurred in the National Needs Served by Educational Measurement?

In the past, success in school was not necessary for entry into the job market. As a result, not all children were expected to succeed in school. Testing practice reinforced the primary goal of education - to select and sort out a small percentage of students who qualified for managerial and professional careers from the majority of students who would ultimately be absorbed into a low-skill workforce. In order to achieve this goal, educational tests that sorted and ranked people gained widespread popularity. The most cost-effective technology for this purpose -- scannable answer sheets and multiple choice test booklets -- was implemented nationwide in schools and colleges, and has become the dominant commercially developed educational measurement tool in use at this time.

Today, the education and assessment paradigms are changing in conjunction with dramatic worldwide power shifts that will impact every aspect of society; particularly the way we educate our citizenry. As unskilled and semiskilled jobs continue to decrease, sorting and selecting lose importance. A higher goal is to provide for each student instructional help so powerful that anyone with adequate motivation can succeed. Finding a way to offer such help to all children is necessary in order to meet President Bush's goal for the year 2000 of providing an education sufficient to meet international workforce demands. Technology may be the only feasible means to provide both instruction and assessment at the level of intensity required.[4]

Our perception of standardized testing must be reconsidered in a new assessment context - one that recognizes the value of thinking, creativity, decision-making, teamwork and the technological and interpersonal skills required to succeed. Conventional tests have merit, but it is what we are measuring and how that measurement is integrated with the professional practice of education that must change. The solution must integrate assessment with instruction and learning. It must reflect clear national goals, both for student achievement and for teaching practice. It requires sensitivity to the economic needs of states, and should celebrate the uniqueness of the individual.

PURPOSES AND OBJECTIVES OF EDUCATIONAL MEASUREMENT AND ASSESSMENT.

A key requirement of the statement of work is to review computer applications to "the various objectives of assessment." A chapter by Millman and Greene in the third edition of "Educational Measurement" provides a good review of purposes and objectives for testing. As a part of a fundamental description of purposes for testing, these authors distinguish between educational measurements taken

[3] Mary Ann Roe, Education and U.S. Competitiveness: The Community College Role. (Austin, TX: IC2 Institute, The University of Texas at Austin, October, 1989). Roe discusses the changes needed in the education continuum, from K-12 and on into the college and workplace to produce a world-class workforce. Beyond the workforce, we need world-class citizens with knowledge and wisdom to vote and uphold the higher accomplishments of our civilization, and with the leisure, means, and desire to pursue service and culture.

[4] W. C. Norris, "The Future of the Information Era.", in R. E. Heldman, (Ed.) Telecommunications Management Planning: ISDN Networks, Products and Services. TAB Books, (Blue Ridge Summit, PA: TAB Books, 1987). pp. 262-263.

BEFORE, AFTER or DURING an educational sequence[5]. In this paper, the concept of *educational sequence* may be viewed as several years of study (e.g., elementary, middle, or secondary school, two- or four-year college programs, professional programs) or as a shorter sequence (i.e., a short sequence of courses, a single course, or a unit within a course). Before the educational sequence a person is an applicant or a recruit. Afterward they are graduates, drop-outs, or failures. In process, they are learners. The current system is designed to provide teaching and resources to help them make progress toward successful graduation from that sequence. Thus, in educational assessment, the following exemplify purposes and objectives before, during, and after.

Purposes of Measurement Before an Educational Sequence

- o For selection among applicants
- o For placement of recruits
- >o For guidance services
  - Course of study planning
  - Learner profiling

Purposes of Measurement During an Evaluation Sequence

- o Incremental grading
- o Incremental failing (quizzes and mid-terms along the way)
- o Routing to non-academic and "special" tracks
- >o Measurement services to help individual students monitor progress
  - extrapolate future progress
- >o Help in formulating the sequence and adapting the sequence to the progress of each learner
- >o Learning feedback: Hints and helps while students grapple with the task
- >o Advice to teachers for grouping students for instruction or for projects
- >o Clarifying the assessment standards and making these assessment standards the explicit goals of learning

Purposes of Measurement After an Educational Sequence

- o Assigning final grades and failures
- o Graduation
- o Certification
- o Licensing
- o Selection for jobs
- o Selection for scholarships, awards, etc.
- >o Guidance services
  - Career guidance
  - Vocational counseling
  - Exit counseling

At the beginning of each list above are bulleted items ("o") that represent assessment imposed externally on individuals and generally used for making high-stakes decisions about their opportunities, rewards, or punishments. The remaining items in each list, designated by ">o", are items that we will call "help services." These are typically referred to as guidance services when given to applicants before or to graduates or failures after. There is a whole class of help services during the process of learning; less common, but these may be far more important in achieving the nation's goals.

Unfortunately, measurement in process during an educational sequence is an area left relatively untouched by testing companies and measurement professionals. It is an area left largely to teachers. This state of affairs represents an anomaly in the field of educational measurement. Leaders call for

[5] Jason Millman and Jennifer Greene, "The Specification and Development of Tests of Achievement and Ability", in R.L. Linn, (Ed.), Educational Measurement, Third Edition, (New York, NY: MacMillan, 1989). (See especially Table 8 on page 336 that summarizes purposes for testing)

assessment that will help improve, not just bring good or bad news[6]. Measurement professionals have long sought the "holy grail" of measurements that truly help learners and teachers, but the dominant professionally developed tests continue to be high-stakes tests given before or after an educational sequence. The productivity of measurement scientists in developing and implementing new measurement instruments that will help teachers and learners in process is disappointing. Most of the attention and resources are given to work on externally imposed measures to support high-stakes decisions. Thus Stiggins was constrained to point out in 1988 that:

> *"Amid the whirlpool of publicity, political turmoil, and scholarly debate currently surrounding the development of a nationally standardized test, or statewide assessments, and of measurement driven instruction, we are again failing to address the central issue in school assessment: insuring the quality and appropriate use of teacher-directed assessments of student achievement used every day in classrooms from coast to coast."*[7]

Teachers develop the best measurement tools they can, but the models they follow derive from the dominant practices they have experienced: assessment with strong consequences before or after educational sequences. They have learned norm-referenced testing, so they write tests that spread students out and then grade on the curve. Often they use grading as a way of providing feedback to learners, but frequently they unwittingly violate vital standards for validity of an assessment: e.g., they assign numbers that do not correspond to what is being measured (such as deducting points from a math test score for tardiness, as though mathematics achievement and punctuality were on the same measurement scale). They make inferences from test scores that are not valid (e.g. giving a low grade on a supposed measure of educational knowledge to punish or control behavior). They may be unaware of the plethora of possibilities for help services opened up by computer technology, which can provide continuous measurement.

Contrasting High-Stakes Assessment and Help Services within the Domain of Individual Assessment

In a high-stakes assessment, a person or group with appropriate authority and professional credentials makes an interpretation of information about an individual that can have a major impact on that individual's life. The term "help services" is probably not a familiar one to most readers of this paper; nor is it a familiar term in the educational measurement literature. It is a term used in this paper to refer to the use of educational assessment to guide and help the learner in accomplishing important educational goals. The before and after uses are familiar ones; guidance and counselling. This guidance may be "high stakes" in the sense used herein, if a critical interpretation and decision is made that affects the person's life, and that decision is made without the examinee's consent, but if the course of action is left up to the individual, it is classified here as a help service.

Table 1 characterizes the distinction between high-stakes assessment and help services for individuals.

Nine points of distinction are given. In each the acronym TEST (Test - Education Sequence - Test) is used to refer to the high-stakes test (usually given before or after the sequence), and the term Help System is used to refer to a help service that uses measurement continuously to improve the process of learning and instruction.

[6] Paul H. O'Neill, Chairman of President Bush's Educational Policy Advisory Committee, and CEO of Aluminum Company of America: "We have some good tests that tell us something, but they don't tell us in a way that would allow us to make specific interventions in the process... They tell us we're not doing very well; they don't suggest why not." Quoted in, Robert Rothman, "2 Groups Laying Plans To Develop National Exams", Education Week, Vol. X, No. 4, (1990 Editorial Projects in Education, Sept. 26, 1990).

[7] Richard J. Stiggins, "Revitalizing Classroom Assessment: The Highest Instructional Priority," Phi Delta Kappan, (January 1988). pp. 363-368.

Table 1
Contrasting High-Stakes Tests vs. Help Systems in Individual Assessment

Characteristics of High-Stakes TESTs	**Characteristics of Help Systems**
Any measurement, whether a TEST or a Help System, is designed to provide relevant, reliable, fair, and timely information to professional people (or to people who aspire to act in a professionally appropriate manner) to help them make valid and defensible decisions.	
1. WHY: TO RANK STUDENTS OR TO ADVISE THEM?	
1. To Rank Them. The BEFORE TEST is designed to spread people out as much as possible along the score scale. A wide score spread facilitates ranking, and thus promotes comparisons between those who are higher and lower on the scale. The AFTER TEST is usually designed to spread people out to facilitate grading, but in criterion-referenced measurement it is used to determine "passing" at some carefully established level.	*1. To Advise Them. For a Help System ranking is irrelevant, and "scores" may not even be visible. What is displayed to the learners and teachers is information to help them make better decisions that will facilitate progress from one step to the next.*
:. WHAT IS THE NATURE OF THE DECISIONS TO BE MADE?	
2. High-stakes decisions about individuals. These decisions can have a significant impact on the future activities and opportunities of the testee, but the decisions are made for them by someone else.	*2. Low Stakes Decisions about Learning Progress. Decisions about which larger units in an educational sequence to take next or about how to correct and improve within a task, have a small impact, and mistaken decisions can be corrected quickly.*
3. WHO ARE THE PROFESSIONAL DECISION MAKERS?	
3. Officers of an Institution. For TESTs the professionals are admissions officers, State and District Administrators holding schools accountable, school psychologists, faculty groups considering graduation requirements, and the like. They seek defensible information to back sometimes unpopular decisions. Teachers make high stakes grading decisions.	*3. Learners and Teachers. It is the teachers themselves, and the more advanced learners who have internalized the standards of excellence, who are the professionals. Psychologists and other professionals cannot stand by teachers and learners to guide them in making valid inferences from every measurement.*
4. HOW DOES TESTING USUALLY TAKE PLACE?	
4. Separate testing sessions, usually obtrusive to the flow of learning and teaching, where printed test materials are distributed and used under strict supervision.	*4. Unobtrusive measurement: The same materials are used to learn or to produce a student product that are used to measure.*

5. WHAT PROFESSIONAL STANDARDS ARE REQUIRED?	
5. Very High. Proper use of TESTs requires professional knowledge and experience. It requires knowledge of professional standards for interpretation and use of tests in general (e.g., that TEST scores should not be the sole basis for a high-stakes decision). It also requires specific knowledge about the particular measurement instrument and the situation. *Proper use of TESTs may be sacrificed for administrative convenience and may be given lower priority than the goal of lower costs.*	*5. High. Proper use of Help Systems will also require professional knowledge, particularly knowledge about a measurement standard -- what it means to be good at something at progressive levels of excellence.* *There will emerge standards for the use of Help System information which, like TEST information, can be misused. For example, it will hopefully become common knowledge that Help System data should not be used in grading students or evaluating teachers.*
6. HOW SECURE MUST BE THE MEASUREMENT TASKS, SCORES, AND REPORTS?	
6. Utmost security must be maintained to assure that the items or tasks, and key, will not be known by any test takers in advance, and that the scores and reports will not fall into the wrong hands.	*6. Free Disclosure. The standards and the form of the tasks should be known in advance. Complex tasks may be practiced as often as necessary before the high-stakes tests. (The learner, however, should have the right to keep progress scores private).*
7. WHEN MUST THE DECISION BE MADE (TIMELINESS)?	
7. Delayed. It is acceptable to have a gap of several weeks or months between the time of TESTing and the time of decision-making. The decision is important and is scheduled in advance.	*7. Rapid, even Immediate. Decisions informed by a Help System are the day-to-day, minute-by-minute, decisions learners and teachers must make. These decisions cannot wait for scoring and interpretive data from a distant location.*
8. WHEN: FREQUENCY OF THE MEASUREMENTS WHICH INFORM THE DECISIONS?	
8. Infrequent. A single TEST, rather than a sequence of measurements, informs most high-stakes decisions (grading is an exception, where a sequence of tests and quizzes is often used).	*8. Continuous. A Help System provides a continuous sequence of measurements, repeated cycles of Teach <--> Test (more accurately, cycles of Practice <--> Coach). Measurement is often indistinguishable.*
9. WHERE DOES SCORING AND REPORTING TAKE PLACE?	
9. It occurs at a central site with fast scanners and paper processing machines, along with large computers and an expert staff.	*9. It takes place in a decentralized educational setting where the data is immediately available to teachers and learners.*

The recommendation that educational assessment practice emphasize help systems in the future gains support from the comparisons in Table 1. If America's needs have shifted from selection and judging to improving achievement for a demographically complex student body, assessment methods are needed that are linked closely with instruction, that are designed to advise and help students and teachers make better decisions about learning progress at the moment of need, are decentralized to the places of learning, and are continuous. As in all assessment practice, high professional standards must be maintained; but different standards will be required, and they will have to be adopted by the teachers and the more advanced learners. Assessment, in contrast to measurement, requires balanced human judgement.

Purposes and Objectives of Program Evaluation

The term program evaluation refers to the collection of data to guide decisions about educational sequences, and aspects of them. By contrast, the term individual assessment refers to assessments made about individual students. The focus of the statement of work for this study was on individual assessment. Program evaluation will be given less emphasis in this paper, and this term will be used to include evaluations of individual classes or special interventions as well as evaluations of the teachers' performances within defined educational sequences. We also use this term to refer to statewide and national assessments, which are used to make interpretive conclusions about the progress and problems revealed by measurements taken using national and statewide samples.

A useful distinction between two forms of program evaluation is summative and formative evaluation. Summative evaluation usually emphasizes measurements taken after an educational sequence, although it may compare them to measures taken before to show gain. The kind of decisions made by summative evaluation are to approve a particular program or terminate it. Formative evaluation, on the other hand, interprets after measures in the light of process measures and looks for ways to improve the process so that the desired outcomes will be achieved.

Summative and formative approaches to program evaluation are contrasted in Table 2.

Table 2
Some Contrasts Between Summative and Formative
Approaches to Program Evaluation

SUMMATIVE PROGRAM EVALUATION	FORMATIVE PROGRAM EVALUATION
1. Why?: *Any program evaluation, whether formative or summative, is designed to provide relevant, reliable, fair, and timely information to the appropriate decision-makers to help them make valid and defensible decisions about programs, and their teachers.*	
2. What Decisions are Characteristic of Program Evaluation?	
2. High-stakes decisions about programs and about key personnel roles within programs. Decisions are made to approve or discontinue.	*2. Decisions are made to refine and improve: to emphasize or de-emphasize particular program components; to allocate resources to revise and improve a component.*
3. Who Makes the Decisions?	
3. Administrators with program authority and budget control.	*3. Developers, teachers, and users concerned with improving the program. Administrators approve funds earmarked for continuing or special improvements.*
4. When are the Decisions Made?	
4. At infrequent intervals when programs are targeted for summative evaluation and review.	*4. Ideally, continuously in cycles of a semester or a year. At a minimum, during the initial trial runs of a program.*

5. How are the Measurements Taken?	
5. The most common summative method is to look at aggregated student scores on standardized tests so that comparisons can be made across different schools.	*5. Fine-grained scoring, both student summaries and direct measures of the system are used. Measuring individual system components makes it possible to highlight specifics that need revision.*
6. Where Do the Evaluations Take Place?	
6. At multiple sites of program operation. General standardized measures are collected that can be summarized across sites. Reports are prepared centrally.	*6. At multiple sites. General program improvements are sought for all, but with localization and customization. Local teachers and users play a greater role.*

Three Measurement Methods for Program Assessment

1. Summarizing and aggregating individual measures. Both completely objective, machine-scorable measures and holistic measures requiring human judgment will be part of a good program assessment.

2. Direct system measures. In a computerized learning and testing environment, direct measures can be obtained, such as the number of lessons passed per student, average time, errors, difficulty of the lessons, and approach to or avoidance of certain elements within the lessons. Task engagement can be measured directly. Do students experience certain kinds of instruction or not? Do they encounter and engage critical instructional elements?

   This is a very promising area for future development. Other direct system measures could include time on task, mean time to help, and effectiveness of each kind of help.

3. Long-term outcomes. Future accomplishments in later classes, higher education, and workplace settings often give a different picture than course grades when included in a program evaluation. Selection test scores and grades in initial academic classes may be highly related to one another, but may not be related to long-term measures, which after all, are much closer to what our nation needs than first-year grades.

A Note on Assessing Teachers. States use multiple choice tests with the flavor of verbal and mathematical minimum competency exams in licensing teachers, weeding out those who fall below a certain cut score. In licensing teachers in this manner, legislators and state administrators have a dilemma. They can show the public that they are striving for higher standards by pushing up the cut score on the tests, but to do often results in the politically unacceptable consequence by rejecting too many prospective teachers from racial and linguistic minorities, who did not do as well as whites on the verbal and mathematical items. If this cut score is forced downward by these political realities, then those accepted into teacher education programs just above this cut score have almost as much need as those below the cut score for remediation and improvement. Could not help systems for the professional growth of teachers, both preservice and in-service, reduce this dilemma by setting higher standards while helping more candidate teachers achieve them? Do not formative evaluation systems offer a hopeful approach when they take direct system measures and focus on how the instruction operates? Could these systems not take some of the focus off of poorly defined qualities of teachers and focus on system attributes that could be changed by management and by better tools?

Summary of Purposes: A Four-Fold Classification of Purposes for Educational Measurement

The distinction between externally imposed measures for high-stakes decisions versus help services is analogous to the distinction between summative evaluation and formative evaluation. Thus, a way to summarize the broad purposes for measurement considered in this paper is to consider individual assessment separately from program evaluation, and consider both High-Stakes and Help purposes for each.

Table 3
Four Purposes for Educational Assessment

Assessment to support:	High-Stakes Assessments	Help Services
Decisions about individuals *INDIVIDUAL ASSESSMENT*	Selection, Placement, Grading, Failing, Graduating, Licensing, Certification, Selection for awards...	Guidance, Course planning, Progress Monitoring, Diagnosis, Advice on strategies, tactics, standards...
Decisions about programs *PROGRAM EVALUATION*	Summative Evaluation to grade and judge educational programs, and teachers, and to hold educators accountable.	Formative Evaluation to identify areas of strength, areas to improve, and ways to improve.

METHODS FOR MEASURING THE ACHIEVEMENT OF INDIVIDUALS

In this section, four methods for educational measurement of individuals will be introduced. These will form the major organizing framework in both Sections II and III. All four are applicable toward any of the four categories of purpose shown in Table 3. The four methods are (1) item tests, (2) standardized performance tasks, (3) student products (exhibits), and (4) process measures taken during tool use.

Item Tests. Item tests are made up of test items of a familiar nature. These usually require very short time intervals to complete, and are appropriate for sampling widely and shallowly from information domains. The most common items are multiple choice items that have only one correct answer. Item tests measure scaffolding knowledge well; memorized terms, facts, and short procedures that can be taught when we "cover" a curriculum rather than teaching it for integration and for transfer to new situations.

Standardized Performance Tasks: These standardized tasks require the integration of multiple lower level pieces of information, and simple skills, in order to perform integrated and complex activities, e.g., solve a problem, design and conduct an experiment, write a document or prepare a presentation or demonstration. The task takes much longer than a simple test item and has integrity, unity, and reference to socially valued roles that students might find interesting and relevant. Performance tasks, unlike items, need not have a single correct answer. They are scored to reflect different paths or solutions. Students may be given partial credit. Holistic assessment is commonly used in grading performance, thus, human judges rate the standardized essay, experiment, documented problem solution, etc. on a scale that may range up to nine or ten points. Different levels of score have different meaning, so that students may be given specific feedback on the standards for any level. Ideally, students should be permitted to repeat the performance task until they are satisfied that they understand the holistic standard for excellence.

Creative Products (Exhibits). An example from athletic competition illustrates the difference between a standardized performance task and a creative student product. The olympic figure skating contestant must first perform a set of standardized "school figures", like figure 8's and jumps. Holistic ratings on a 10 point scale are used, and the ratings of several judges are averaged. Then the contestant presents a free-style program. It must meet certain criteria, but the composition, choreography, amount of risk, etc. are up to the contestant. The judges have agreed in advance to a set of standards to rate the free-style programs on the same 10-point scale. The contestants fully understand the criteria used in the rating system.

Examples of student exhibits are writing assignments, reports, presentations, performances, designs, artistic productions, etc. The tasks cannot be standardized, or else the room for creativity is diminished. This sort of objective goes beyond integration to enable the assessment of transfer of learned knowledge and skill to a new situation, putting together what has been learned in a new way; adding new insights not directly presented in the way the material was originally learned.

The result of the student's creative production is an exhibit, which must also be scored holistically for primary traits agreed to in advance. In ideal situations, these primary traits are fully understood by both teacher and student. Like the Olympic contestant, they know the difference between a performance that is rated 8, 9 or 10. Moreover, the more completely the learners understand the holistic scoring standard, the more valid that measure becomes. This is in contrast to coaching for an item test. With an item test, the more we focus on narrow objectives and structured types of questions, the less valid the test becomes. Frederiksen and Collins,[8] who introduced these authors to the Olympic skating example, advocate holistic scoring of student exhibits as a method of increasing "systemic validity;" that is, the more you teach to it, the more valid it becomes, whereas when you teach to item tests, you teach tricks for guessing and things to memorize to avoid thinking. This makes the test less valid for measuring the desired level of cognitive functioning.

Indirect Measures During Tool Use: Suppose a student uses an outline processor, followed by a word processor, for writing an essay in response to a creative production assignment. Whenever the student uses one of these software productivity tools, he or she is interacting with the computer and the responses can be scored and interpreted. Research and development is needed, and intelligent software, to perform this interpretation and to generate hints and helpful advice to help improve the student strategies.

It is not necessary to wait until intelligent scoring programs can be put on-line during tool use. Both standardized performance tasks and student exhibits may be utilized now by using holistic scoring schemes performed by human raters. As teachers and students learn to assign these holistic scores, they achieve an important educational objective of understanding in a deep fashion what it means to be excellent. The use of computer tools during the development of a creative product facilitates this process. The ability to print out intermediate products, as with a word processor, and discuss and evaluate them is an excellent way to integrate assessment with instruction aimed at creative production objectives.

Examples of learning strategies are library research tasks that require searching strategies, interviewing and questioning techniques, note-taking and review techniques, and methods for controlling emotional states such as anxiety and procrastination. Examples of performance strategies are found within both academic and athletic games, and within tasks like writing, speaking, or performing.

SUMMARY

Section I dealt with the evolving purposes, objectives, and roles of educational assessment in the 1990's. Measurement and assessment were shown to be fundamental to all occupations and

[8] John R. Frederiksen and Allan Collins, "A Systems Approach to Educational Testing", Educational Researcher, Vol. 18, No. 9 (Washington, DC: American Educational Research Association, December 1989). pp. 27-32.

professions because decision-making is fundamental to all, and decision-making depends on sensitive judgments based on appropriate and accurate information.

Educational measurement plays a peculiar and skewed role compared to measurement in other professions. There is a lack of consensus on what should be measured, and measurement practice is skewed toward using the measures for judging, grading, sorting, and selecting. The current measurement technologies have grown out of the successful use of aptitude tests for sorting and selecting. However, grading practices and testing for accountability in achievement measurement resemble these aptitude testing practices too strongly. Thus, a case was made for the development of a new family of measurement applications called help systems more integrated with instruction, and geared toward achievement through learning progress for a demographically diverse student population. Such a development will better serve current national needs, which are poorly served by finding better methods for selecting and passing judgments.

Program evaluation, including the assessment of teachers, also appears to be skewed toward judging more than helping. It could profit from much greater focus on measurement integrated with instruction for providing help to improve the teachers and their programs.

Section I introduced a distinction between measurement and assessment. Measurement is a vital process of deciding what attributes to measure, and then providing accurate data so that decision-makers can determine presence or absence, and more or less of the attributes selected. Educational assessment uses measures, <u>and</u> other information, about individuals or programs to make an interpretation and then a decision. Assessment requires human judgment that goes beyond the accurate measurement of some attribute.

Human judgment is required in three of the four measurement methods discussed in section II.

1. Objectively scored item tests
2. Holistically scored standardized performance tasks
3. Holistically scored student products (exhibits)
4. Human judgments about intermediate products in the process of developing a student exhibit on performing a task.

None of these measurement methods are new. Standardized performance tasks have been a part of individually administered tests for intelligence and clinical diagnosis for many decades, but they require highly trained professionals to administer the tasks, rate the responses, and interpret the resulting profiles. Both standardized written assignments and assignments for creative writing are common, but also require holistic grading. Good teachers of writing, speaking, presenting, acting, performing, athletic coaching, and crafts judge intermediate products or performances and provide hints or helps along the way.

How have computerized methods been used in administering each of the four measurement methods? Can computerized methods go beyond substituting or incrementally improving item tests for high-stakes purposes, and beyond the conventional uses of item tests for selecting and judging? Can computers transform assessment by partially automating the scoring of standardized tasks, student products, and by providing hints and helps during process? The latter contribution would be transformational. In addition, the computer's role in introducing a much closer integration between instruction and assessment would be transformational. How much promise do computers have for transformational applications?

In the next section, these and other issues will be addressed. We will show that computers have so far been used primarily to substitute for and incrementally improve item tests. But this paper will show that computerized educational assessment offers great promise for introducing the help systems and the formative evaluation systems that will enable teachers to develop professionally in enhanced roles as assessors and managers; students to achieve higher, more complex achievement objectives than are measured by conventional tests, and educational systems to evolve through measurement and formative evaluation toward the kinds of productive systems our nation needs.

SECTION II: CURRENT USES OF COMPUTERS IN ASSESSMENT

TYPES OF COMPUTERIZED MEASUREMENT SYSTEMS

The term Computerized Educational Assessment System (CEA System) will be used in this paper to refer to a system that uses computers in some part of the administration process for an educational measurement instrument. Computerized systems for administering tests can be grouped into six categories, depending upon the mode of administration used. Al'. of the six CEA System types use computers in the processes of scoring and reporting of results. The six CEA Systems are:

Portable, Non-Interactive Answer Media
1) Scanned answer sheet systems
2) Portable keypad systems
3) Bar code readers

Interactive Testing Stations
4) Computer work stations
   - o Learning stations
     - > in Classroom (Group Dicplay)
     - > in Cluster (Individual or Small Group Use)
   - o Specialized lab work stations
5) Customized simulator environments
6) Specialized notebook computer systems

The six CEA System types are grouped into two categories: Portable Non-Interactive Answer Media and Interactive Testing Stations. Answer sheet systems do not use computers to present the displays or accept the responses, but do use computers to scan the answer sheets, score them, and print out reports. This widely used testing technology exists because of computer technology and paper handling technologies. Furthermore, its portability is one of its important assets, since testing can occur in any room with the suitable desks or tables. Portable keypad systems and barcode readers that do not interact can substitute for the answer sheets and the scanners. The responses go directly into a digital form without the intermediate scanning step. The barcode readers are currently being used to aid in the process of holistic scoring of student essays. Many national and state testing programs now incorporate student essays which must be graded by two or more human assessors. The bar-code readers help manage the data and aid in rapid decisions to add a third reader if the first two ratings diverge widely.

Most computer workstations and customized simulation environments are not portable, so a new infrastructure must be set up consisting of local rooms equipped with a sufficient number of computer workstations or simulators. Schools have learning centers with workstations that can be used for instruction or for measurement. Interconnection to a central file server is desirable to collect the records from each student's testing session.

The notebook computer system is so small and portable that it has the potential to enable interactive testing to compete with scanned answer sheets. When fully developed for computerized measurement, such systems would also replace the portable keypad systems and barcode readers.

Essential Processes in Administering an Educational Measurement Instrument

Considering the six processes of test administration listed below reveals some of the variations among the six types of CEA systems. Not included are other processes measurement experts use in developing, distributing, and in statistical analysis and record-keeping following administration. These are discussed briefly later in this section.

1] Presenting Item or Task Displays: Three methods are in use, portable displays for each individual (e.g. printed test booklets), interactive computer displays used one-on-one, and group displays using a variety of media.

2] Obtaining a Record of Responses: The most common method is to use the scannable answer sheet. It is most often used in an individually paced mode. That is, the students are given a certain amount of time to work a certain number of items that they can sequence in any manner and arrange their time as they wish. When the items are presented in a group mode, for example, when an auditory listening test is given in language assessment, students are all given the same amount of time to complete each item. Interactive computers in labs or learning centers are individually paced and use a variety of response entry devices (keyboard, mouse, joystick, touch screen, voice pickup). Notebook computers are portable and would lend themselves either to individual pacing or to group pacing.

3] Scoring and reporting: Getting the responses into the computer and providing a scoring key in the computer is the challenge. Scanners for answer sheets have made mass group testing possible. They can and should be improved to go beyond multiple choice items. Developing a scoring key for complex performance tasks, even when all responses are collected by computer, is no easy task. Once scored, software for generating a variety of reports is available and is quite mature.

4] Interpretation of results for individuals: Conventional answer sheet systems are scored objectively and require no interpretation during the scoring. This has a cost advantage and removes the bias from subjective rating, since human graders are not immune to bias or to differing interpretations. Interpretation of the results comes after the answer sheet is sent back to be scanned, scored, and the results printed out. On some clinical tests, computerized tools are now in use to interpret profiles of scores from psychological instruments that produce a profile. The best of these are "expert systems" that have captured the rules expert assessors use to interpret profiles.

Individually administered tests of intelligence and of psychological diagnosis require substantial human assessment throughout. When tests are individually administered, the expert human administrator judges and scores each performance for each task. Sometimes the score is objectively determined, but more often the student's vocalizations or movements must be interpreted.

5] Monitoring testing sessions: Human proctors or trained teachers preside over testing sessions to assure standardization and fairness, and to deter cheating. Computer methods have been used in computerized testing centers: video cameras, time control, alarm lights, etc. Monitoring for overall test security goes beyond session monitoring. The security of test booklets and answer keys must be maintained during development and distribution, as well. Electronic distribution provides new means, such as encryption, to solve this old security problem in high-stakes testing. It also provides new risks because it offers those so inclined new ways to gain access to the items and the keys, and the possibility for tampering with the scores.

6] Special practices for special populations: Handicapped people require special provisions. They may be given extra time. The visually impaired may require a human reader. Computers, i.e. educational testing stations, can be equipped with special response devices (audio, braille keyboards, unconventional response devices for physically handicapped people, etc.). Willingham[9] has published a comprehensive volume on issues in testing handicapped people with conventional tests.

More work is needed to integrate the creative engineering for computer interfaces for handicapped people into assessment practices. Interactive computer displays can blow up printed information, provide headphones and volume controls for audio, and can provide a variety of special response devices for those who are physically handicapped and cannot use a keyboard or mouse. Computers can also control time intervals very precisely, and it is possible that a much more equitable way to adjust the timing for handicapped people can be determined. For example, it takes blind students longer to read a passage in Braille than sighted students take to read the text. Perhaps the computer could determine when the student had finished reading and then provide equal time for answering. By extension, it might be possible to provide different amounts of time equitably for test takers from different language groups. Even as Braille is more time-consuming than English text, reading an English text is more time-consuming for a

[9] Warren Willingham et al., Testing Handicapped People, (Boston, MA: Allyn and Bacon, Inc., 1988). These authors considered students who were learning disabled: hearing impaired, visually impaired, and physically handicapped. Their work was sponsored by the College Board Educational Testing Service, and their Graduate Records Examinations Board, so the research was conducted with paper and pencil and multiple choice and standardized essay tests.

native Spanish-speaker than for a native English-speaker.

ADVANTAGES AND DIFFICULTIES OF ANSWER SHEET SYSTEMS

Advantages

Interactive testing stations have several challenges if they are to compete favorably with the existing predominant use of answer sheet systems. Consider the following strengths of answer sheet systems and item tests.

1. Relatively low cost administration, scoring, and reporting. Many gradations of scanning and processing equipment are available, so each application can find a cost-effective implementation. Processing larger volumes of tests offers economies of scale.

2. Portability. Usability in many locations and settings.

3. Widespread familiarity and acceptance. The public is familiar with test booklets and answer sheets and accepts them with equanimity. Studies with new item types that require people to think, rather than eliminate and, if necessary, guess, reveal that thinking items are not welcome. Students agree that open-ended items that do not give the answer among a set of distractors are "probably more valid". Still, they find them much harder, more time consuming, and they dislike them.

4. Coverage. Because each item takes such a short time, much more content and variations in cognitive demand can be covered in the time available for testing.

5. Reliability. Related to the larger number of items that can be taken in a given time period, the scores are more reliable than with tests consisting of fewer items.

6. Predictive Validity. The admissions tests predict first-year grades.

7. An existing infrastructure, both technological and human. From central sites to the network of part-time test administrators, conventional tests have a well-established infrastructure that is not in danger of being supplanted soon. Capitalizing new hardware for each school is not enough. It is also necessary to train the larger group of test developers, test statisticians, teachers and other users how to use the equipment and software effectively.

Criticisms of Item Test/Answer Sheet Systems

Among the many criticism of the common testing format are:

1) That our tests rely on many brief and unconnected items, and thus measure only the temporary existence of snippets of knowledge. Instead, students need to develop an integrated personal knowledge structure that can provide the learner with an organized, powerful, generative and deep form of knowledge useful in adapting to changing circumstances.

2) That our tests measure less important outcomes and miss many things of obvious importance to success in an increasingly technological society -- such things as critical thinking and reasoning, problem-solving, trouble-shooting, flexibility, creativity, motivation, persistence, and strategies for learning and self-control.

3) That the dominant objective item formats, in particular, multiple- choice promote strategies that have nothing to do with important educational outcomes -- strategies of elimination, guessing, time allocation.

4) That focussed preparation for a broad collection of snippets of knowledge tested via specific test formats takes time that should be spent on attaining the deep and powerful knowledge and thinking skills needed to be a contributor after schooling.

5) That the use of narrow tests of minimums focusses instruction and squeezes out the more difficult but desirable outcomes.

6) That the tests are artificial, indirect, and not given in an interesting, integrated, real-world context.

7) That the tests are biased toward certain ethnic and gender groups.

8) That the items have but one correct answer, portraying education as a search for a collection of little right answers, rather than as a process of forming one's own questions and evaluating several partially right alternatives to complex problems.

Minimum competency testing, considered in the next section, has come in for special criticism.

Minimum Competency Testing and Its Effects on What, How and Whom We Teach

Varying federal, state and district policies for mandated achievement testing using standardized norm-referenced and criterion-referenced tests have tended to *narrow the curriculum*, encouraging teachers to *teach to the test*. As a result, classroom teachers devote a significant portion of time in test preparation and practice testing prior to the mandated state and/or national tests. This narrowing of the curriculum has focussed classroom instruction on an outdated notion of basic skills (basic reading and arithmetic) that has left little time for teaching science and other complex subjects, nor time for advanced skills, higher-order thinking, reasoning, and problem solving skills.

Clearly, there is a direct relationship between what we test and what we teach. Two maxims concerning assessment programs summarize the relationship succinctly: "What you test is what you get." and "What you don't test you don't get."[10] Mehrens and Kaminiski have written concerning the issue of how closely teachers should teach to a test.[11] They discuss a continuum of teaching to the test and recommend that the point where one crosses from appropriate to inappropriate methods depends on the inferences the teachers wishes to make from the test scores. If the teacher is interested in inferring to a larger domain of knowledge, then it is inappropriate to teach narrowly and specifically to a sample of scaffolding objectives drawn from that larger domain.

Standardized item tests have also significantly effected how we teach (and how we think about teaching). Since standardized items are generally quite short, and have one and only one correct answer, we teach students that knowledge can be broken down into short, simple additive components. We may

[10] Lauren B. Resnick, "Tests as Standards of Achievement in Schools," Paper presented at the 1989 Educational Testing Service Invitational Conference Proceedings, The Uses of Standardized Tests in American Education, New York, October 28, 1989

Lauren B. Resnick and D. P. Resnick, "Assessing the Thinking Curriculum: New Tools for Educational Reform," In Bernard R. Gifford and M. C. O'Connors (Eds.) Future Assessments: Changing Views of Aptitude, Achievement, and Instruction (Boston, MA: Kluwer Academic Publishers, in press).

Lorrie A. Shepard, "Why we need better assessments", Educational Leadership, April, 1989, pp. 4-9.

[11] William A. Mehrens and John Kaminiski, "Methods for Improving Standardized Test Scores: Fruitful, Fruitless, or Fraudulent?", Educational Measurement: Issues and Practice, Spring, 1989, pp. 14-30.

not teach students that the solution and problem solving path we use is as important as the answer. Likewise, we have not taught students that there may be several alternative correct answers. In our attempts to cover the broad scope of material addressed in the state curriculum plans, we have not spent sufficient time on the "powerful ideas" and the "core, essential concepts" of the disciplines. If achievement tests are the operational goal, they offer a way to select and narrow to a smaller target.

We need to teach a smaller number of powerful ideas well, rather than a broad coverage of content without a coherent structure. Furthermore, we have tended to view the content disciplines of mathematics, reading, writing, and science as separate and independent knowledge domains. We do not teach students to understand the relationships within and among these knowledge domains. We typically do not have students write or speak about their own mathematical ideas, or read original source materials about great mathematicians.

Finally, by advocating the use of standardized, item-based tests, the minimum competency movement has influenced whom we teach. For example, results on standardized achievement tests are generally used as a primary indicator for retaining students at their current grade level rather than providing appropriate instruction and remediation so these students can progress to the next grade with their peers.

A recent three-year study by the National Commission on Testing and Public Policy, entitled "From Gatekeeper to Gateway: Transforming Testing in America[12]," charged that the American testing system has become a "hostile gatekeeper" which limits opportunities for many students, particularly women and minorities. The commission called for a innovative, transformed assessment system which would "open the gates of opportunity for America's diverse people."

Superintendents often demand that principals and teachers "raise test scores" as the most important goal. It is not emphasized that the test scores are only proxies or indicator behaviors for the real learning outcomes which we expect from schools.

As Shepard[13] points out, teachers may direct students away from good instruction when using a standardized test to identify mild handicaps. The results from these tests significantly harm students by labelling them. The label becomes the explanation for the observed behavior, "He cannot read because he is learning disabled". Then they are redirected into less challenging classes, with lower expectations, and where there is less teacher encouragement and pressure for learning progress. Shepard also discusses the effects of errors of measurement. In one study nearly half of the students labelled as learning disabled were really normal or were average performing students in above-average performing classes or schools.

SUBSTITUTIVE, INCREMENTAL, AND TRANSFORMATIONAL APPLICATIONS OF INTERACTIVE TESTING STATIONS

Three Stages in Technology Diffusion

The U.S. Office of Technology Assessment has found that the diffusion of a technology goes through three typical phases. These phases are:

1) <u>The substitution phase:</u> The newest technology is used as a more efficient substitute for the older manual or labor-intensive procedures. For example, the first applications of computers to assessment were simple computerized tests that duplicated the exact items in the exact item sequence, and used the same scoring procedures of their printed test equivalents.
2) <u>The stage of incremental improvements:</u> Working with the implementation of technology at the substitution level, inventive people soon discover incremental improvements that utilize features of the technology not used in its substitution phase. Examples will be given of tests where items

[12] National Commission on Testing and Public Policy, <u>From Gatekeeper to Gateway: Transforming Testing in America</u>, (Chestnut Hill, MA: National Commission on Testing and Public Policy, 1990).

[13] Lorrie A. Shepard, "Identification of Mild Handicaps" In Robert L. Linn (Ed.) <u>Educational Measurement, Third Edition</u> (New York, NY: Macmillan Publishing Company, 1989) pp. 545-572.

are selected dynamically according to a mathematical model. Another incremental improvement still making an impact is the move from the monochromatic printed content of most test items to increased use of computer graphics, color and even animation in the item contents.

3) The introduction of new approaches to accomplishing fundamental goals: Inherent features of the technology that were dormant during the substitution and incremental improvement phases are introduced in this phase. The goal of the original activity is reconceptualized at a deeper level.

Use of Interactive Testing Stations to Administer Item Tests

The second edition of Educational Measurement (1971) includes some early references to the potential of testing by computer[14]. These original computerized testing applications typically employed mainframe and minicomputer systems accessed by computer terminals. Widespread use of computers for testing was significantly spurred by the advent of the integrated circuit and microprocessor in 1970, followed by the first personal microcomputers in the late 1970's.

Typically, computerized testing involves the conversion or translation of conventional paper-and-pencil tests to a computer-administered format with little or no change in the basic test structure or procedures.

Table 5 highlights the major benefits and limitations of computerized tests compared with conventional paper-and-pencil tests.

Table 5
Benefits and Limitations of Computerized Testing

Technology Benefits

- o Greater standardization of test administration
- o Immediate test presentation
- o Immediate test scoring and reporting
- o Enriched display and response capabilities
- o Allows new item types and item formats
- o Reductions of certain types of measurement error
- o Ability to measure response latency for items and components
- o Improved capabilities for score analysis and interpretation
- o Improvements in test security
- o Easy aggregation of testing records tests.
- o Creation of customized tests and items by computer

Technology Limitations

- o Limited number of computer terminals per school
- o Incompatible computer hardware and software
- o Need for equating studies between computerized and paper-administered tests
- o Limited computer experience of some students
- o Need to examine equity, bias, and legal issues
- o Possibility of new types of measurement error from computerized testing
- o Lack of imaginative item types when multiple choice format is copied from paper

[14] Frank B. Baker, "Automation of Test Scoring, Reporting, and Analysis", in Robert L. Thorndike (Ed.) Educational Measurement, Second Edition (Washington, DC: American Council on Education, 1971).

Research reviews have examined comparability of test scores from computerized tests and paper and pencil tests.[15] These studies typically show no significant differences or only slight test score differences in favor of one or the other testing mode. These differences are of little practical significance. Recent reviews hypothesize that these small mean score differences may be due to specific user characteristics of a small portion of examinees.[16] These user characteristics might affect performance negatively on computer-administered tests more than the paper-administered tests for a small portion of examinees.

Examples of Computerized Tests

The computer administered conventional test (CT) is the most widespread of the types of computerized assessments. Computerized tests are employed to measure generalized achievement, to diagnose skills and learning capabilities, personality characteristics, learning aptitudes, and mastery of instructional objectives.

An illustrative computerized test is the WICAT Comprehensive Assessment Test.[17] This computerized testing product includes comprehensive tests of reading, mathematics, and language arts for grades K-8. The computerized tests measure a common set of educational objectives addressed by all of the major standardized achievement tests. A testing management system is provided for selecting students and tests, for sequencing the tests, administering the tests, and generating computerized testing reports. The computerized test items include text, graphics, and digitized voice quality audio. Directions for taking the test are given with text, graphics, and digitized voice-quality audio.

Educational Testing Service has developed computerized versions of two College-Level Examination Program (CLEP) tests and is currently conducting research to verify comparability of scores from the computerized and paper-and-pencil tests. One of the CLEP tests employs digitized, photographic illustrations to test artistic judgment.[18] Educational Testing Service has also developed interactive assessment videodisc demonstration projects for medical certification and English as a Second Language.[19]

District-wide implementations of computerized testing have been demonstrated for measuring state assessment objectives.[20] Widespread use of computerized professional certification tests has also been

[15] C. Victor Bunderson et al., "The Four Generations of Computerized Educational Measurement," In Robert Linn (Ed.) Educational Measurement, Third Edition (New York, McMillan Publishing Company, 1989).

John Mazzeo and Anne L. Harvey "The Equivalence of Scores from Automated and Conventional Versions of Educational and Psychological Tests: A Review of the Literature," Research Report No. CBR 87-8, ETS RR 88-21. (Princeton, NJ: Educational Testing Service).

[16] Steven L. Wise and Barbara S. Plake, "Research on the Effects of Administering Tests via Computer," Educational Measurement: Issues and Practices, vol. 8, no. 3, Fall 1989, pp. 5-10.

[17] Wicat Systems. Wicat Comprehensive Assessment Test. (Wicat Systems, Orem, UT, 1990).

[18] William C. Ward, "ETS Innovations in Assessment," (Princeton, NJ: Educational Testing Service, 1990).

[19] B. Bridgeman, Randy Bennett, and S. Swinton, Design of an Interactive Assessment Videodisc Demonstration Project (Princeton, NJ: Educational Testing Service, 1986).

[20] James B. Olsen, Arryl Cox, Charles Price, Mike Strozeski, and Idolina Vela , "Development, Implementation, and Validation of a Computerized Test for Statewide Assessment," Educational Measurement: Issues and Practice, vol. 9, no. 2, Summer 1990, pp. 7-10.

demonstrated.[21]

Computer-Based Psychological Testing and Interpretation

An emerging area of computerized testing research and implementation is computerized psychological tests and interpretations. Computerized testing versions have been developed for several psychological tests such as the Minnesota Multiphasic Personality Inventory, Ohio Vocational Interest Survey, Self-Directed Search, Functional Skills Screening Inventory, Marital Adjustment Test, and cognitive style tests.

Computer Based Test Interpretations (CBTI) are also provided for several psychological, personality, and vocational related tests. Computerized test interpretations provide detailed text and graphics based reports which interpret the results of psychological tests. Professional guidelines have been adopted by the American Psychological Association concerning computer based psychological tests and test interpretations.[22]

INCREMENTAL IMPROVEMENTS IN COMPUTER-ADMINISTERED TESTS.

1. Computerized Adaptive Testing

Reckase notes that, "Adaptive testing, the dynamic selection of items to match the performance of an examinee during the administration of a test, has finally become a readily accessible methodology for use in standardized testing programs[23]." This form of testing has achieved a milestone with the publication of a comprehensive "Primer"[24] that presents in terms as accessible to lay readers as possible the significant benefits and strong technical foundations behind this important innovation in testing.

A computerized adaptive test is a computerized test in which the next item or task is adaptive or tailored depending on the examinee's previous responses. In a computerized adaptive test, an item of average difficulty is administered first. If the examinee answers the item correctly, a more difficult item is presented. If the examinee answers the item incorrectly, a less difficult item is presented. The adaptive testing process continues until a specified stopping rule is reached and the testing process terminates. Typical adaptive test termination criteria include a fixed number of test items, a minimum standard error, or a maximum information value.

Computerized adaptive testing is based on pioneering developments in item response theory.[25]

Dean A. Slawson, District-wide Computerized Assessment in Texas (Provo, UT: Waterford Testing Center, 1986).

[21] Allan C. Bugbee, Jr., "Students Prefer Computer Administered Testing," Educational Measurement: Issues and Practice, vol. 8, no. 4, Winter 1989, p. 28.

[22] American Psychological Association, Committee on Professional Standards and Committee on Psychological Tests and Assessment, Guidelines for Computer Based Tests and Interpretations (Washington, D.C.: American Psychological Association, 1986).

[23] Mark D. Reckase, "Adaptive Testing: The Evolution of a Good Idea," Educational Measurement: Issues and Practices, vol. 8, no. 3, Fall 1989, pp. 11-16.

[24] Howard Wainer, (Ed.) Computerized Adaptive Testing: A Primer Erlbaum Assoc., Hillsdale, N.J., 1990.

[25] Fred M. Lord, Applications of Item Response Theory to Practical Testing Problems (Hillsdale, NJ: Lawrence Erlbaum Associates, 1980).

Ronald K. Hambleton, "Principles and Selected Applications of Item Response

Computerized adaptive tests include four major components: a pool of test items from which the test is created, a procedure for selecting items from the pool, a method for computing the test score when the test is completed, and a means for determining when the testing should be terminated. Additional information on the components and standards for computerized adaptive tests have been documented.[26]

Computerized adaptive tests yield all of the benefits presented above for computerized tests. In addition, computerized adaptive tests also provide the following benefits and limitations presented in Table 10.

Table 6
Benefits and Limitations of Computerized Adaptive Tests

Technology Benefits

- o Increased measurement precision with significantly fewer items
- o Tests/items are individually selected accordingly to each examinee's responses
- o Increased testing efficiency with time savings of 50% to 70%
- o Rapid and accurate measures at all ability levels
- o Improvements in test security
- o Ideal for ranking and grading. Spreads people out along a single dimension.

Technology Limitations

- o Limited number of computer terminals in schools
- o Requires large numbers of student responses for item calibration and analysis
- o Unidimensional item response and scaling model does not necessarily reflect stages of complex cognitive growth
- o Schools are not structured to take advantage of the time savings from adaptive tests
- o Requires more advanced test interpretation skills

Examples of Computerized Adaptive Testing. One representative example of computerized adaptive testing is the College Board Computerized Placement Tests developed jointly by the College Board and Educational Testing Service.[27] The Computerized Placement Tests are computerized adaptive tests designed for use by two- and four-year colleges to assess if entering students are ready for college level work in English, reading, and mathematics, or need additional developmental courses. These tests have been used for a period of four to five years at approximately 80 colleges across the U.S.

An additional example of computerized adaptive testing is the Differential Aptitude Tests, Computerized Adaptive Edition published by Psychological Corporation.[28] This test battery provides

Theory" In Robert L. Linn (ed.) Educational Measurement, Third Edition (New York, NY: Macmillan Publishing Company, 1989) pp. 147-200.

Ronald K. Hambleton and Hariharan Swaminathan, Item Response Theory: Principles and Applications (Boston, MA: Kluwer Academic Publishers, 1985).

[26] Bert F. Green, Darrell B. Bock, Lloyd G. Humphreys, Robert Linn, and Mark D. Reckase, "Technical Guidelines for Assessing Computerized Adaptive Tests," Journal of Educational Measurement, vol. 21, no. 4, Winter 1984, pp. 347-360.

[27] L. J. Abernathy, (1986) Computerized Placement Tests: A Revolution in Testing Instruments. (New York, NY: College Board).

[28] The Psychological Corporation (1986). Differential Aptitude Tests, Computerized Adaptive Testing Edition. (San Antonio, TX: The Psychological Corporation, Harcourt, Brace Jovanovich, Inc.)

League for Innovation in the Community Colleges, Computerized Adaptive Testing:

computerized adaptive tests for the following aptitudes: verbal reasoning, numerical ability, abstract reasoning, clerical speed and accuracy, space relations, spelling, and language usage. This computerized adaptive testing battery is available in both IBM PC and Apple II versions. The test is used in junior and senior high schools.

With the emergence of microcomputers, computerized adaptive testing has now become feasible for widespread operation and implementation, as well as for research. Within the past few years, a wide variety of computerized adaptive testing systems have been developed, demonstrated and implemented by organizations including: American Institutes for Research, American College Testing Program, Assessment Systems Corporation, Educational Testing Service, Psychological Corporation, and Wicat Systems. Computerized adaptive testing applications are presently available for achievement testing, aptitude testing, compensatory education selection tests, college entrance and placement tests, professional certification, and licensure tests.[29] Computerized adaptive tests have also been developed for district and statewide assessment.[30]

2. Computerized Mastery Tests

In comparison to computerized adaptive testing, which attempts to obtain accurate measurement across a broad range of proficiency levels, computerized mastery tests seek to provide accurate measurement at the cut score or decision point which separates masters from non-masters. The computerized mastery test presents items which help to discriminate examinees above and below the mastery cut score. Computerized mastery testing is the preferred model of choice for most certification and licensing programs. The theory and procedures for computerized mastery tests have been documented.[31]

Examples of Computerized Mastery Tests

Educational Testing Service has developed a computerized mastery test for the National Council of

The State of the Art in Assessment at Three Community Colleges. (Laguna Hills, CA: League for Innovation in Community Colleges, 1988).

[29] Mark D. Reckase, "Adaptive Testing: The Evolution of a Good Idea," Educational Measurement: Issues and Practice, vol. 8, no. 3, Fall 1989, pp. 11-16.

Susan Grist, Lawrence Rudner, and Lauress Wise, " Computerized Adaptive Tests" ERIC Clearinghouse on Tests, Measurement, and Evaluation, Digest no. 107 (Washington, D. C.: American Institutes for Research, February, 1989).

[30] G. Gage Kingsbury, "Adapting Adaptive Testing with the MicroCAT Testing System" Educational Measurement: Issues and Practice, vol. 9, no. 2, Summer 1990, pp. 3-6.

Sue M. Legg, Dianne Buhr, and Robert Wickham, "Adaptive Testing for State-Wide Assessment, MicroCAT News, March 1989, pp. 1,4,5.

Jose Stevenson, "Computerized Adaptive Testing in the Montgomery County, Maryland Public Schools," MicroCAT News, April 1987, pp. 1, 4.

[31] Fred M. Lord, Applications of Item Response Theory to Practical Testing Problems. (Hillsdale, NJ: Educational Testing Service, 1980).

David J. Weiss and G. Gage Kingsbury, "Application of Computerized Adaptive Testing to Educational Problems" Journal of Educational Measurement, Vol. 21, 361-375.

Architectural Registration Boards.[32] Test questions are organized into a series of short structured testlets designed to match the overall test content specifications and to provide equivalent measurement characteristics. It would be inappropriate to select individual items randomly from a pool or even according to difficulty level as in a computerized adaptive test. To do so would almost always violate the rules for content, coverage and balance found in the test's specification. Testlets also serve to correct unexpected context effects; for example, if the computer selects two items from a pool, the first item might give away the answer to the second one inadvertently[33]. A minimum number of testlets are drawn randomly from the available pool of testlets and then administered to the examinee. At the conclusion of each testlet, a decision is made concerning whether or not the examinee should be classified as a master or non-master based on performance from the combined testlets. Computerized mastery tests typically require only half of the questions administered in the conventional paper-and-pencil format.

3. Computer-Based Diagnostic Testing. A particularly intriguing application of computerized testing for educational purposes is the computer-based diagnostic test.[34] A computer-based diagnostic test attempts to identify the specific conceptual, procedural, or performance errors which the student makes in response to test items or testing situations. Some diagnostic tests attempt to diagnose and classify cognitive errors within a generalized problem solving domain. These errors are often referred to as misconceptions or cognitive "bugs."

4. Incremental Improvements to the Display of Item Contents

Computerized Video, Graphics and Animation Tests. The increasing capability of microcomputers to display still frame video, motion video, high resolution graphics, and animations provide for increasingly more realistic and challenging types of test items. These capabilities provide for assessment of interactive and dynamic characteristics similar to real life situations. With video and photographic display capabilities of microcomputers, educators can develop and administer tests of science, social studies, art, history, and languages which are very realistic and life-like. How much better would be a science test which included photographs, motion segments, and high resolution animated color graphic displays of science concepts and processes?

Examples of Video, Graphics and Animation Tests. In 1979 the National Science Foundation funded

[32] William C. Ward, "ETS Innovations in Assessment," (Princeton, NJ: Educational Testing Service, 1990).

[33] Howard Wainer and G. L. Keily, "Item clusters and computerized adaptive testing: A case for testlets", Journal of Educational Measurement, 24, (1987). pp. 185-201.

[34] Kikumi K. Tatsuoka, Diagnosing Cognitive Errors: Statistical Pattern Classification Recognition Approach. (Urbana, IL: University of Illinois, Computer-Based Education Research Lab, 1985).

David L. McArthur, Diagnostic Testing Project. (Los Angeles, CA: University of California at Los Angeles, Center for the Study of Evaluation, 1985).

Garlie A. Forehand and Myrtle W. Rice, Diagnostic Assessment in Instruction. Machine Mediated Learning, Vol. 2, No. 4, 1988, pp. 287-296.

Isaac I. Bejar, Educational Diagnostic Assessment. Journal of Educational Measurement, Vol. 21, No. 2, 1984, pp. 175-189.

a proof of concept study for a computer-controlled videodisc addressing college-level developmental biology.[35] This videodisc included computerized testing components, some with motion video segments (i.e., unraveling of the DNA molecule), still-frame video displays, animation graphics, and high resolution graphic display items. Evaluations of the videodisc tests showed that students effectively learned and retained the information presented in motion video, video still frame, and animation displays.

A computerized graphics and animation test has been developed to test science process skills of variable identification, hypothesis formation, operational definition, experimental design, and interpretation of data. The test demonstrated high reliability, and difficulty and discrimination indices which were acceptable for evaluating criterion referenced achievement.[36] A computerized animation test has been developed for a three-dimensional spatial rotation task. The test included 80 three-dimensional rotation items created from eight basic graphic figures.[37]

POTENTIALLY TRANSFORMATIONAL APPLICATIONS OF COMPUTERS

1. Current Uses of Computers to Administer Standardized Performance Tasks

An extensive, two-semester, fifteen-unit Physical Science course has been developed by the Texas Learning Technologies Group[38]. It is in use in eleven other states besides Texas. In the minimum configuration, videodisc or computer displays are presented by the teacher on a monitor at the front of the class. Students work in small groups of four or five around videodisc equipped computers at the back of some classrooms or in a learning center. These same computers can be used for individualized tutorials for individual students. The Texas Physical Sciences curriculum includes a variety of simulations, but these are not scored as a part of the assessment. It is an interesting commentary on the state of the art in scoring standardized performance tasks that Educational Testing Service was approached to assist TLTG in developing the assessments for this innovative curriculum and developed a set of multiple choice paper and pencil administered tests (ETS is involved in other projects involving the scoring of simulation tasks).

Recent nationwide trends in educational assessment favor the use of performance-based assessments as alternatives to the traditional multiple-choice standardized tests.[39] Performance tasks require students to publicly display and effectively use their personal knowledge and skills to write, discuss, think, solve complex problems, and conduct experiments. Examples of performance tasks considered by states

[35] Bunderson, C.V., Baillio, B., Olsen, J.B., Lipson, J.L., and Fisher, K.M. Instructional effectiveness of an intelligent videodisc in biology. Machine-Mediated Learning, 1,2, 1984.

[36] Michael E. Hale, Development of a Computer Animated Science Process Skills Test, Paper Presented at the Annual Meeting of the National Association for Research in Science Teaching (New Orleans, LA: April, 1984).

[37] Isaac I. Bejar, A Psychometric Analysis of a Three-Dimensional Spatial Task. (Princeton, NJ: Educational Testing Service, 1986).

[38] Borich, Gary D. "Outcome Evaluation Report of the TLTG Physical Science Curriculum, 1988-89," The University of Texas at Austin.

[39] Doug A. Archbald and Fred M. Newmann, Beyond Standardized Testing: Assessing Authentic Academic Achievement in the Secondary School (Reston, VA: National Association of Secondary School Principals, 1988).

Grant Wiggins, "A True Test: Toward More Authentic and Equitable Assessment," Phi Delta Kappan, May 1989, pp. 703-713.

Grant Wiggins, "Teaching to the (Authentic) Test," Educational Leadership, vol. 46, no. 7, April 1988, pp. 41-47.

involve direct writing assessments, open-ended mathematics and reading items, integrated reading and writing exercises, and hands-on science experiments.

Results from a recent survey show that nearly half of the nation's states are developing, or plan to develop, performance tasks as a significant component of their statewide assessments.[40] The states of California, Connecticut, Massachusetts, New York, and Vermont are currently implementing statewide performance assessments. The states of Alaska, Arizona, Colorado, Florida, Hawaii, Kentucky, Maine, Missouri, New Jersey, North Carolina, Oregon, and Pennsylvania are currently developing statewide performance measurements. Additional information on performance measurement systems can be found in the following references.[41] Performance tasks have recently received national educational support and interest from a coalition of three dozen educational and civil rights groups.[42]

The most widely employed performance task is the direct writing assessment using standardized essay prompts, currently used in the National Assessment of Educational Progress, the General Educational Development Testing Service, twenty-eight statewide assessments, and in the College Board Advanced Placement Tests. The writing performance task requires students to write a brief essay(s) in response to a specific writing prompt(s). Current computer measurement technology has been applied to the direct writing tests in the following areas:

- o Banks of writing prompts,
- o Word processors as alternatives for students to use in creating the written essays,
- o Text data bases and editors which teachers can use for storing, retrieving, and managing the student essays.
- o Barcode readers for recording holistic writing scores

Future computer technology developments are expected to provide additional capabilities for measuring student writing performance including: computerized handwriting recognition systems, automated handwriting and text conversion systems and automated scoring of student essays.

Examples of Performance Tasks. Educational Testing Service has just announced a computerized portion of the National Teacher Examinations, the most widely used teacher licensing exam.[43] In the

[40] Pamela R. Aschbacher research described in Robert Rothman "New Tests Based on Performance Raise Questions" Education Week September 12, 1990, p. 1,10.

[41] Lauren B. Resnick, Education and Learning to Think (Washington, D.C.: National Research Council, 1987).

Doug A. Archbald and Fred M. Newmann, Beyond Standardized Testing: Assessing Authentic Academic Achievement in The Secondary School (Reston, VA: National Association of Secondary School Principals, 1988).

Joan B. Baron, "Blurring the Edges Among Assessment, Curriculum and Instruction," Paper presented at the Education Commission for the States and Colorado Department of Education Assessment Conference, Boulder, CO, June 1990.

Grant Wiggins, "A True Test: Toward More Authentic and Equitable Assessment," Phi Delta Kappan, May 1989, pp. 703-713.

Grant Wiggins, "Teaching to the (Authentic) Test," Educational Leadership, vol. 46, no. 7, April 1988, pp. 41-47.

[42] Campaign for Genuine Accountability, "Statement on Genuine Accountability," Education Week, January 31, 1990, pp. 1, 12.

[43] Karen Diegmuller, "E.T.S. Previews Revamped Examination for Teachers" Education Week, Vol. 10, No. 2, September 12, 1990, pp. 1, 13.

computerized test, the teacher candidates will be asked to supply their own answers as well as select from multiple-choice responses. The teachers will also be asked to write brief computerized essays. Adaptive testing features are also included.

The National Bureau of Medical Examiners and the National Council of State Boards of Nursing are currently developing and pilot testing computerized adaptive performance tests for professional certification of physicians and nurses. These tests include applied patient management problems with multiple correct answers, open ended performance items, and medical simulation exercises.

The military and commercial training industries have also developed a wide variety of performance tasks and work simulators. Many of these work-based performance tasks and simulations employ computer technology for developing, administering, scoring, and reporting. These computerized performance tasks range from the combat games and two dimensional flight simulators which run on single or networked personal computers to the complex three- dimensional, full-flight simulations which are used to train airline pilots and flight technicians.

2. Current Uses of Student Products in Assessment

Increased national interest in performance testing has also lead to an emphasis on student exhibits and portfolio methods.[44] A student portfolio includes a representative collection of the student's work over a sustained period of time requiring concentrated effort and providing a perspective on personal progress the student has made toward exemplary performances. For example, the variety and range of student exhibits in a mathematics portfolio might include: written journals, biographies of investigation, student conference presentations, student designs and inventions, investigative reports, physical or computer mathematical models, videotapes, and/or reflective essays.

Interactive computer technology provides a very effective vehicle for assisting students in the development and management of creative products, exhibits, and portfolios and assisting teachers in the evaluation of these creative performances, exhibits, and portfolios.

3. Process Measures During Tool Use

Using integrated desktop software systems such as the Apple MacIntosh or Microsoft Windows 3.0, it is feasible to collect computerized process measurements as students select, use and integrate each of the available electronic educational tools. These process measures can be used to evaluate student time spent with various tools, sequences and patterns of tool use, most frequent activities with each tool, and generalized learning and problem solving strategies. The collection, analysis, and reporting of process measures during tool use will require a generalized instructional management system which is able to collect data nonintrusively.

The future holds promising potentials for unobtrusive measures and feedback during and after the use of tools in the production of exhibits, presentations, etc. Currently, tools like word processors are stimulating instructional valuable dialogue between teachers, students, and fellow students as intermediate products are reviewed and discussed. The ease of changing drafts in the computer promotes experience and instruction on review and revision.

4. Computer Applications That Integrate Assessment With Instruction

Increasingly, testing specialists and educators are strongly recommending the need to integrate assessment with instruction. In his preface to the third edition of Educational Measurement, Robert Linn

[44] Dennie P. Wolf, "Portfolio Assessment: Sampling Student Work," Educational Leadership, vol 46, no. 7, April 1989, pp. 35-40.

Dennie P. Wolf, "Opening Up Assessment," Educational Leadership, April 1988, pp. 24-29.

reiterates the historical need for integration of assessment with instruction.[45]

> *"In my view, the biggest and most important single challenge for educational measurement today is no different from what it was at the time the first edition of this book appeared; that is, to make measurement do a better job of facilitating learning for all individuals. However, to date, measurement has done a much better job of predicting who will achieve and of describing that achievement than of helping teachers adapt instruction to enhance the learning of individual students. The combined efforts of cognitive psychologists, measurement specialists, and educators will need to be devoted to this task if educational measurement is going to become, not "a process quite apart from instruction, but an integral part of it".*[46]

Linn and Tyler are only two of the many measurement professionals who have written concerning the need to integrate assessment with instruction.[47] Two of the primary methods currently employed for integrating assessment with instruction include Computer Managed Instruction and Integrated Learning Systems.

Computer Managed Instruction. Computer managed instruction (CMI) systems use the computer to record and manage much of the routine data associated with managing an entire classroom, school, or district in which the students are working at differing instructional levels, with different curriculum materials, and with differing achievement levels. A computer managed instruction system typically consists of a bank of instructional objectives, a large item bank, lesson pretests, curriculum materials and exercises, lesson post-tests, and a bank of instructional prescriptions.[48] Item banks are used to create the required pre- and post-tests. CMI tests in the past were usually administered in paper-and-pencil format with a computer-readable answer sheet. The answer sheets were scanned using desktop or high-speed scanners or the score results are entered by teachers using a keyboard. When the tests are given

[45] Robert L. Linn, "Current Perspectives and Future Directions." In Robert L. Linn (Ed.) Educational Measurement, Third Edition (New York, NY: McMillan Publishing Company, 1989).

[46] Ralph W. Tyler, "The Functions of Measurement in Improving Instruction" In E. F. Lindquist (ed.) Educational Measurement (Washington D.C.: American Council on Education, 1951) p. 47.

[47] Frank B. Baker, "Technology and Testing: State of the Art and Trends for the Future," Journal of Educational Measurement, vol. 21, no. 4, Winter 1984, pp. 399-406.

Nancy S. Cole, "Future Directions for Educational Achievement and Ability Testing," In Barbara S. Plake and John C. Witt (eds.) The Future of Testing (Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers, 1986).

Garlie A. Forehand and C. Victor Bunderson, Basic Concepts of Mastery Assessment Systems (Princeton, NJ: Educational Testing Service, 1987a)

Michael E. Martinez and Joseph I. Lipson, "Assessment for Learning," Educational Leadership, vol 46, no. 7, April 1989, pp. 73-76.

Anthony J. Nitko, "Designing Tests Integrated With Instruction," In Robert L. Linn (ed.) Educational Measurement, Third Edition (New York, NY: McMillan Publishing Company, 1989) pp. 447-474.

Lorrie A. Shepard, "Why We Need Better Assessments," Educational Leadership, vol. 46, no. 7, April 1989, pp. 4-9.

[48] Frank B. Baker, Computer-Managed Instruction: Theory and Practice (Englewood Cliffs, NJ: Educational Technology Publications, 1978).

on learning workstations, the term CMI is not usually used. CMI systems may include networks among several personal computers and several desktop scanning machines. The mainframe-based CMI systems such as Project PLAN, Individually Guided Education, and the Navy CMI systems, implemented in the mid 1960's and 1970's, promised to provide individualized education through widespread use of printed curriculum materials and computer-scannable answer sheets. Each of these large-scale CMI systems has been retired. Current CMI system implementations typically employ microcomputers linked through modems and local area networks.

Integrated Learning Systems (ILS) were designed as integrated learning environments for students, combining assessment, instruction, and management within a single system. An ILS typically includes the following hardware components: a central file server, mass data storage devices, local area communications network, thirty or more personal computer workstations, and a printer. The primary software components include: an instructional management system, comprehensive computer assisted instruction, computerized testing and assessment software, staff development activities, and instruction-related software tools.

The ILS environment provides students with opportunities to take computerized achievement tests and receive appropriate prescriptions for mastered and non-mastered objectives. The prescriptions include comprehensive computer assisted instructional modules designed to teach non-mastered curriculum content and objectives. Student responses to the courseware lessons and computerized testing materials are monitored and teachers receive reports on individual students and class performance as students proceed through the curriculum lessons or computerized tests at their own pace.

USE OF COMPUTERS IN PROGRAM EVALUATION

Computers play a critical role in collecting, analyzing and reporting data from local and national evaluations of many educational programs (i.e. Chapter 1 and Chapter 2 of the Elementary and Secondary School Act, National Assessment of Educational Progress, National Educational Longitudinal Study, etc.).[49] Program evaluation is conducted primarily to judge the worth and value of educational programs. Educational programs are judged as valuable if they lead to significant improvements in student learning as measured by student achievement test scores and other educational indicator variables. Data from individual student and group pre-test scores are compared with data from student and group post-test scores to determine if an educational program produces any significant achievement gains or losses. Computer data base management systems and statistical packages are often used to manage the program evaluation data and to conduct statistical analyses of the individual and aggregated group achievement results. National evaluations of educational programs also use computers to conduct nationwide statistical analyses of educational program effects and outcomes.

Computer applications for program evaluation are generally found at the large district, state and national levels. However, to the authors' knowledge, there is not a uniform national system or computerized network for use of computers for program evaluation. If each district had a comparable computer and the program evaluation data from each district and aggregate the results at each state and from each state to the national educational agencies. New computer applications should also be developed for assistance in district and state needs assessment, review of proposed evaluation designs and instruments, and statistical analysis, and database management of the program evaluation data.

Use of Computers for Aggregating Individual Data

Computers provide ideal data collection, aggregation, analysis and reporting tools. With local area networking and long haul telecommunications capability, data concerning certain elements of student achievement can be aggregated from the classroom to the school administration office, from the school administration office to the district administration office, from the district administration office to the state educational office, and from the state educational office to the national Department of Education. Data

[49] Herbert J. Walberg, and Geneva D. Haertel, The International Encyclopedia of Educational Evaluation (New York: Pergamon Press, 1990).

base management systems can be used at each level of aggregation to store, query, retrieve, and report results from various subsets of the program evaluation data. Statistical packages can also be used at each appropriate level of aggregation to provide statistical analysis of educational program effects. Several states (Florida, Ohio) and districts (Azusa, CA; Pharr San Juan, and Alamo, TX; and Anne Arundel County, MD) are currently implementing district-wide and statewide information networking systems which will facilitate aggregation and integration of student achievement data.

Use of Direct System Measures

The increased availability of networked personal computers and integrated learning systems provide the foundation for development of direct measures of the effectiveness, efficiency and productivity of the educational system. The computerized learning system can record each interaction the student has with the computerized instructional and assessment system. These data are stored in a data base management system for easy access, statistical analysis, and creation of customized or standard reports. Data from student interactions can be used to calculate the following preliminary list of direct measures of system performance: time on task, mean time to help, lesson and response duration, lesson effectiveness evaluation, attrition, lesson avoidance, and estimated completion times. Additional direct system performance variables should be hypothesized and investigated.

Use of Long Term Educational Outcome Measures

The nation relies heavily on standardized item tests with items which can be answered quickly, with one and only one correct answer, and which are generally independent and unrelated to other test items. These are efficient and far less costly than searching for long term educational outcome measures. Measures of success in employment rates after schooling, and in productive accomplishments (e.g., publications, patents) is difficult and costly to obtain. Emphasis has therefore been placed on short term variables which might improve student scores on the multiple choice tests of scaffolding objectives. Thus, we have focused on improvements in one type of learning indicator and have neglected exploration of long term improvements in learning.

The primary long term outcomes expected from K-12 education include basic skill competencies in reading, writing, language, and mathematics for functional real world and job-related contexts; higher order thinking, reasoning, creativity, and problem solving skills; and an increasing repertoire of alternative learning strategies. Long term outcomes should emphasize applied capabilities and performance required for later classes, higher education, and future workplace settings. Long term outcomes should emphasize the need for lifelong learning and education.

To assist in providing closer links between the worlds of school and work, the U.S. Secretary of Labor has established the Secretary's Commission on Achieving Necessary Skills (SCANS Commission).[50] The focus of the commission is to identify essential job-related skills for effective work performance. The initial list includes twenty-eight functional job skills in the areas of resource management, information management, social interaction, systems behavior and performance, human and technology interaction, and affective skills. It is expected that by age 16, all students will have gained proficiencies in these workforce readiness competencies. These competencies provide the potential foundation for measurement of long term educational outcomes.

To illustrate the need for focusing on long term educational outcomes, consider the results from a long term research project with the Graduate Record Advanced Chemistry test. The project[51] found a strong negative correlation between examinees scores on the Advanced Chemistry test and the number of

[50] Michael Kane, Sue Berryman, David Goslin, and Ann Meltzer, The Secretary's Commission on Achieving Necessary Skills (Washington, D. C.: Pelavin Associates, Inc, 1990)

[51] "Long-Term Validity of the Advanced Chemistry Examination", ETS Research Report (1989)

subsequent research papers and publications produced by the examinees. Graduate schools should consider whether they value efficient test-taking students more highly than those who emphasize research and publication.

These data suggest the need to explore the variables which influence long term educational outcomes. Our typical short term perspective on educational outcomes may mitigate against what we really want to measure.

CURRENT USES OF COMPUTERS IN DEVELOPMENT, DISTRIBUTION, AND ANALYSIS OF EDUCATIONAL ASSESSMENTS

Since the mid 1960's computer technology has been employed in selective areas of testing and assessment. Test publishers have used mainframe or minicomputers to enhance productivity in the tasks of test construction, item banking, test printing and test distribution. Word processors are employed for item writing, editing, and review. Item banking programs are used to store large collections of items for easy access to item displays and item characteristics. Additionally, computers are used to facilitate test construction, editing, and review. After the test is developed, laser and color printers facilitate test printing and formatting.

Computerized tools for content, job, and task analysis would be helpful to further define and develop more complex, integrated and motivating assessment and performance tasks. There exists a long standing separation between the content and processes taught in school and the content and processes required in the world of work. To reduce this gap, editable versions of SCANS commission functional skills and assessment scenarios could be made available to state departments and school districts. The states and districts could then customize these components to their own local needs and requirements.

The benefits and limitations of computer uses in test development and reporting are presented in Table 7.

Table 7

Benefits and Limitations of Computers in Test Development and Reporting

Technology Benefits

- o Word processors used for item writing
- o Item banking programs for item search, selection, and insertion
- o Automated test construction, editing, and review
- o Item analysis and calibration
- o Improved test printing and formatting
- o Increased flexibility and ease of test
- o Automated ordering and distribution processes
- o Remote electronic registration
- o Improved test reporting

Technology Limitations

- o Limited number of integrated test construction systems
- o Limited graphics editors
- o Lack of professional item interchange formats
- o Limited computer experience of test developers

Computer technologies have also been used effectively in test registration and distribution processes. Using a touch telephone, computer access for remote registration and scheduling can easily be accomplished. The electronic registration information can be used to schedule the number of test administration sessions and to electronically download the test from a mainframe computer location to distributed testing center loca ions. These same procedures can be used to download tests from mainframe computers to personal computers.

Examples of Computer Uses in Test Development

Excellent summaries of computer uses in test development, distribution, and reporting are provided in the following references.[52]

Test Analysis, Record Keeping and Reporting

High speed test answer sheet scanning machines scan and process answer sheets, score the tests, and store the information in a computer readable format. Large mainframe or minicomputers are then used to process and analyze the testing information and to prepare printed reports for the individual students and groups tested. These mainframe and minicomputers are typically located at centralized test development, publication, and scoring service centers. Test publishers have used computer technologies to enhance their productivity in test construction, item banking, test printing, and test processing and reporting.

Computers are used in test analysis, record keeping, and reporting because they provide for automation of time consuming and tedious human labor tasks. The computer can read, score, and store each of the item responses. Item analysis and item response theory statistics can be calculated easily, and the item and test statistic files can be automatically updated using only a few simple commands. Archival copies of test scores can also be easily made. Computers provide for a wide range of individual and group reports to be printed from the resulting test scores and profiles. Computerized interpretative reports have also been prepared for an increasing number of educational and psychological tests.

THE INFRASTRUCTURE FOR EDUCATIONAL MEASUREMENT AND ITS PROBABLE EVOLUTION

The term "infrastructure" is used in this paper to refer both to the technological delivery system and the human expertise. The educational feeders for sustaining the flow of expertise are also part of the human infrastructure. For conventional paper and pencil testing, the infrastructure is in place. Computerized educational assessment requires the introduction of a new decentralized infrastructure. It

[52] Frank B. Baker, "Computer Technology in Test Construction and Processing," In Robert L. Linn (ed.) Educational Measurement, Third Edition (New York, NY: McMillan Publishing Company, 1989), pp. 409-428.

Tse-Chi Hsu, "Developments in Microcomputer Applications to Testing," Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 1986.

Jason Millman and Jennifer Greene, "The Specification and Development of Tests of Achievement and Ability" In Robert L. Linn (ed.) Educational Measurement, Third Edition (New York, NY: Macmillan Publishing Company, 1989), pp. 335-366.

Jason Millman and Judith A. Arter, "Issues in Item Banking," Journal of Educational Measurement, Vol. 21, No. 4, Winter 1984, pp. 315-330.

Anthony J. Nitko and Tse-Chi Hsu, "A Comprehensive Microcomputer System for Classroom Testing," Journal of Educational Measurement, Vol. 21, No. 4, Winter 1984, pp. 377-390.

Gale H. Roid, "Item Writing and Item Banking by Microcomputer: An Update," Educational Measurement: Issues and Practice, Vol. 8, no. 3, Fall 1989, pp. 17-20.

requires permanent learning and assessment centers at schools and colleges. The mix and configuration of these centers between separate lab rooms and computerized classrooms has not been determined.

Three different groupings of professionals currently make up the human infrastructure that is driving innovation and implementation of the precursors of CEA. These three are:

1. Educational measurement professionals, and the current testing industry.
2. Instructional technologists, trainers, and human resource development professionals.
3. Computer technologists and users from many fields of endeavor.

Each of these has its own strengths and weaknesses. The testing industry, which is both sustained and criticized by measurement scientists and related professionals, will receive the most attention in this review.

A Critique of the Educational Measurement Infrastructure and the Current Testing Industry

We have quoted the words of leaders in the educational measurement profession and have provided footnotes to document that these professionals are very concerned with the need to integrate assessment with instructions, but have not found the resources to shift very much of their research and development into the new priorities, nor have they introduced many new products.

These professionals work at universities, school systems, at some industrial corporations, at the twc major nonprofit centers; Educational Testing Service of Princeton, NJ, and The American College Testing Program in Iowa, and at for-profit testing companies. They constitute the current U.S. educational measurement infrastructure for research and development. Whether they distribute testing products or not, people in many organizations form a professional community with an important collective wisdom for dealing with the complex issues of educational measurement. However, the tendency for this professional community to communicate to its in-group in mathematical and statistical terms tends to isolate it[53]. An adversarial relationship aimed at the testing companies has developed from a growing number of politically active consumer groups. These advocacy groups have gained publicity by attacking different uses of standardized testing without demonstrating a grasp of the complexity of the assessment issues, or proposing sound alternative measurement solutions.

Despite its contribution to the science and technology of measurement, the U.S. educational measurement infrastructure has not kept pace with trends and practices in the other two professional communities. As a result, instructional technologists are more likely to lead out in the development of new applications that integrate assessment with instruction. Neither have many measurement professionals kept up with computer and information technology. These trends have the potential to transform the kinds of human-computer interactions possible in education as well as the outcomes that can potentially be assessed. Thus, as the media for learning is transformed by advances in computing, our definition of literacy, education, and assessment must reflect this transformation.

Current educational tests are defined in relation to interactions with printed matter, rather than in relation to dynamic interactive environments rich in visual and auditory displays, new response formats, or large computerized archives of information. Computer users from many disciplines are developing simulations, visualizations, and interactions that make new kinds of assessment (such as Help Systems) possible. These innovators, however, are not cognizant of educational measurement issues and practices. The need is great to link knowledge acquired from interactive multimedia research with testing and instructional development practice.

How Much Leadership Will the Testing Industry Provide? Professional testing companies have important strengths, but are the target of much criticism. They are national resources of expertise in measurement science, and often ascend to statesmanlike leadership on education issues where their expertise is strong. They promote and try to abide by high professional standards for quality and fairness. Where they maintain significant numbers and quality of researchers, they enlighten the national debate

[53] The publications of NAEP, in general, and of the new ETS policy information center have a refreshing accessibility.

with factual data. They also may pioneer new item types, new computerized methods, even new ways to integrate assessment with instruction.

Various anti-testing advocacy groups attempt to protray them as villians, but their villainy is often illusory. The tests bring bad news about the educational progress of minority groups versus white males, and this is taken as proof of bias in the tests. It may only indicate failure to provide the educational help in homes and schools that children from each group need.

A more valid criticism results from constraints these companies operate under. It is not a conscious policy on their parts, but is very real. Professional testing companies exert a strong conservative and inertial force in slowing the pace of acceptance and implementation of computerized testing and unfamiliar instructionally-oriented systems. Whether organized as for-prophet or not-for-profit, these companies make their income from selling paper test booklets, paper answer sheets, and in scanning and reporting paper-and-pencil multiple choice tests. Companies are thus very reluctant to adopt computerized testing; they are also reluctant to be the first to announce a computerized standardized achievement test. Several companies have been willing to invest in research and development in computerized testing, and have developed some computerized testing products for item banking, localized test scoring, and aptitude testing. However, in the absence of a proven market and delivery infrastructures, none of the major professional test publishers have been willing to announce development or release of a computerized standardized achievement test.

Professional testing companies continue to rely on mainframe and minicomputer technology for test development, research and reporting tasks. In general, these organizations have not emphasized building expertise in microcomputer technology, or in innovative display and response technologies.

Several forces act on testing companies to keep them locked into this conservative posture.

1. Testing companies serve clients, not end users. States, professional organizations, and membership organizations mediate between testing companies and the test-taking public. As with any business, the client is much in control of what types of new testing instruments are developed and deployed. Clients can lead out, and when they do, the testing companies are responsive. The innovative National Council of Architects Review Board is sponsoring R&D that is potentially very significant. The College Board and the GRE board have sponsored some forward-looking research in CEA. The College Board is distributing computerized placement tests, and innovative microcomputer-based testing systems.

An increasing number of states, districts and Canadian provinces have expressed interest in computerized testing. Testing companies will be responsive to these clients, but technology companies may persuade these clients that they can provide a faster, less expensive solution. The statistical and scientific quality standards that the testing companies adhere to are hard to explain; hard to sell.

2. Standards for validation, quality, equating and fairness slow down innovative projects. Professional testing companies should be commended because of their continuing strong commitment to professional test development research and validation standards and to standards of fairness and quality. However, these same commitments to professional standards requiring exhaustive research and validation tend to repress innovation, creative solutions, and exploration with the use of new technologies for testing. It will simply cost too much before it can begin to yield a return on investment.

The idea of formative research -- starting with a partial system and evolving it over time based on field experience, is fraught with too much risk to companies who are judged by the unchallengeable quality of each product as it comes out the door.

3. Testing companies are very concerned about legal challenges. They have had to fight many challenges to the use of their products in certain high-stakes areas (employment selection is perhaps the most hotly litigious). They have not perceived that low-stakes products like help system are so fundamentally different that legal liabilities may be minute.

4. Testing companies do not want to be accused of developing instruction for their own tests. The makers of a high-stakes test is in an awkward position if they also develop the instructional help products to prepare people for those tests.

5. Competitive pressures and project pressures prevent organizations from changing. The test developers are not given time by their managers to try out new item types or delivery options. They must make their quotas of items, or else schedules and budgets will not be met. The innovative developer who takes time off to work on a research project, even if funded from another part of the company, may not be promoted as readily as those who keep fully occupied on bread and butter tasks. These tasks are to develop dependable, accepted and valid multiple choice tests of the highest quality the world has ever known.

This observation is not unique to testing companies. They are business organizations, albeit with high ideals. They must meet their client's schedules and produce an income to survive and thrive.

6. The capital investment required is too high for them to risk. The capital investment in new modes of testing means high costs for R&D that must be diverted from improving the bread-and-butter printed tests, high costs for restructuring the company internally and retooling people expert in paper processing to become good at computers, and an expensive, missionary-type of selling to convince people to install hardware with the features of ILS systems in order to run the new tests. Such investments are questionable, to say the least, for a supposed market that does not have an infrastructure in place.

6. Measurements professionals have an academic mistrust of business. The more academically oriented a testing company is (and the non-profits tend to be quite academic), the less comfortable they feel about embarking on new business strategies that involve high amounts of capital investment and risk. Hardware for permanent testing centers, and a new technically literate human infrastructure is very costly.

The Probable Evolution of Current Testing Products

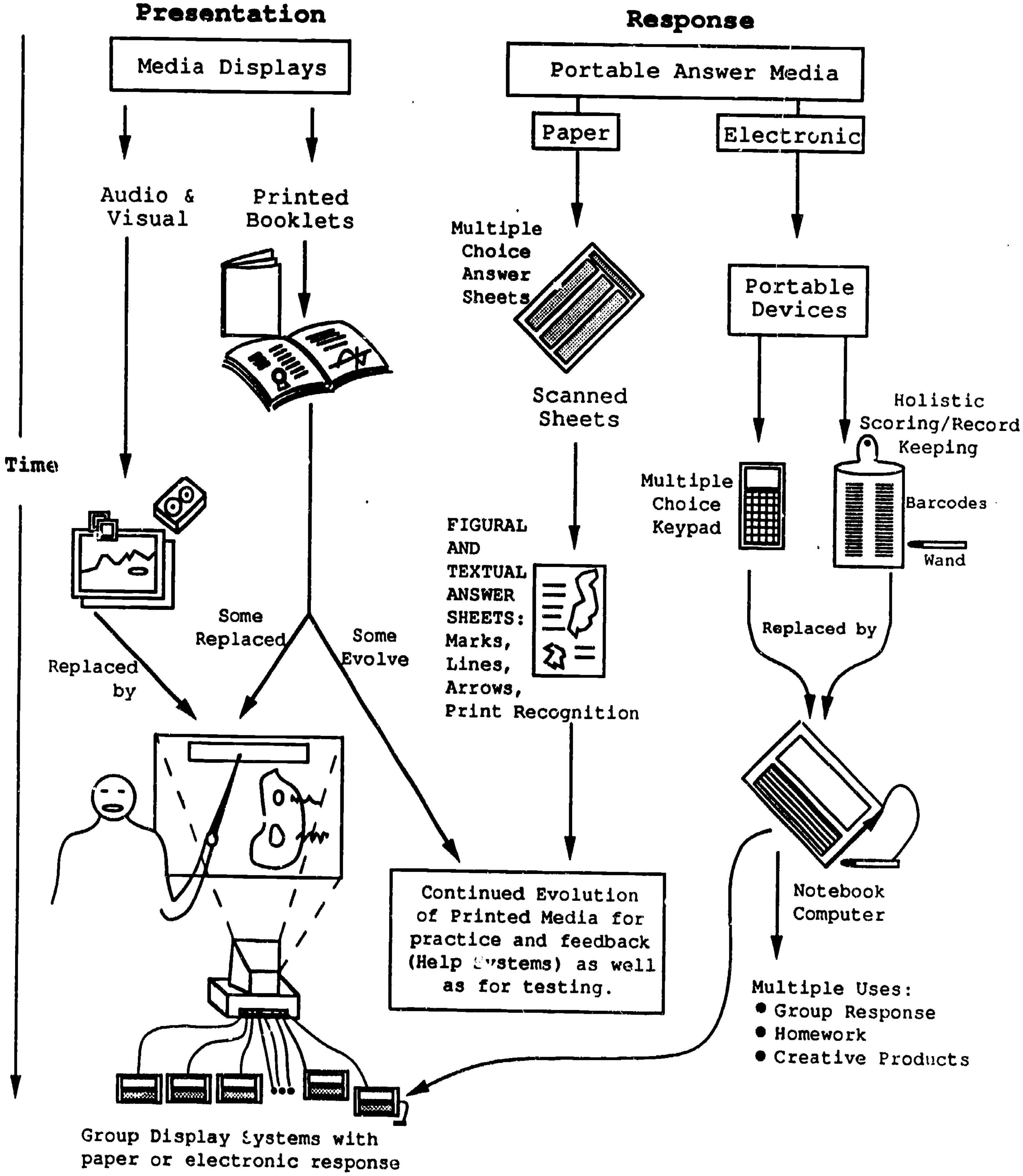
Figure 1 shows some possible evolutionary progressions in the delivery of answer sheet/item test measurement instruments. A test requires stimulus presentations, and response entries correlated to the displays. Printed booklets and answer sheets dominate, but some tests use audio-visual media. Listening and language tests require audio or video tapes, or their equivalents. These group testing modes (individual response, group pacing) will give way to group presentation systems with either answer sheets or electronic response devices. Help systems (practice with feedback worksheets) can be implemented with paper, although computer delivery has far more benefits, but at greater costs and loss of portability. Notebook computer-like devices offer both the functionality and the portability of printed tests.

Probable Evolution of Answer Sheets. Answer sheet systems should evolve and will do so. Because of the established infrastructure, both technological and human, and the low cost, the portability, and the public familiarity with these systems, it is desirable that actions be taken to promote the continued evolution of answer sheet systems. It is probable that answer sheet systems will continue to evolve as scanners provide higher resolution and as new item types beyond multiple choice are developed to utilize these systems. Test answer sheets are divided into several hundred "bubbles" -- small ovals or rectangles where the scanner looks for a mark. Most test developers use this type of sheet only for multiple choice, but it is possible to use an answer sheet for other item types. For example, by providing a grid for each item in a math test of 4 columns and 13 rows of bubbles, the rows respectively designated by the symbols "-"(minus), "/"(divide), "."(decimal), or a digit from 0 to 9, students could enter arithmetic expressions such as 9/15, 62.8, 45-7, and many others into this grid without selecting from 4 or 5 multiple choice alternatives[54].

[54] Brent Bridgeman, A comparison of multiple-choice and free-response quantitative questions on the graduate record examination. ETS Research Report (in press). Bridgeman took away the distractors in graduate record exam math items, and presented the same item stems with grids to 3000 students. The "grid-in" items are much harder than the original multiple choice items, and were superior in measuring at the high end of ability. The erroneous responses entered by most students were not usually the distractors invented by the test

Figure 1

EVOLUTION OF ANSWER SHEET/ITEM TEST SYSTEMS


Presentation
Response
Media Displays
Portable Answer Media
Paper
Electronic
Audio & Visual
Printed Booklets
Multiple Choice Answer Sheets
Portable Devices
Scanned Sheets
Holistic Scoring/Record Keeping
Time
Multiple Choice Keypad
Barcodes
Wand
FIGURAL AND TEXTUAL ANSWER SHEETS: Marks, Lines, Arrows, Print Recognition
Some Replaced
Some Evolve
Replaced by
Replaced by
Continued Evolution of Printed Media for practice and feedback (Help Systems) as well as for testing.
Notebook Computer
Multiple Uses:
● Group Response
● Homework
● Creative Products
Group Display Systems with paper or electronic response

It is also possible to arrange paragraphs so that the words, printed in non-scannable ink, each fall across several bubbles. Then words can be marked or lined out by the student and these selections can be scored[55]. Recent research at ETS has resulted in a new family of "Figural Response Items"[56]. These items utilize high-resolution optical scanners that pick up tiny picture elements (pixels) from lines and marks drawn by the student. The computer reconstitutes the locations and configurations of these pixels and compares them with specifications in a scoring key. The figural response items presented to the student contain pictures on the answer sheet. The student marks a part of the picture (e.g., mark the nucleus in the cell), draws an arrow (e.g., which way will the ball travel when it emerges from the curved tube?), or draws a line (e.g., where could you cut the flatworm to yield the pictured cross-section?).

The art of handwriting recognition is developing, and will eventually mature to the point where printed letters and numbers can be recognized from scanned answer sheets. Figure 1 depicts these developments.

Answer Sheet Systems in the Classroom. Answer sheet systems can certainly participate in the transformation of learning and instruction through integrating assessment with instruction. Teachers can use computers to display both instruction and assessment information to whole classes. Students can mark their responses on answer sheets, which could be scanned on an optical scanner attached either to the teacher's computer or to a multi-purpose computer at the back of each classroom. The measurement activity could be quickly scored and could be made a part of a timely group discussion.

Possible Leadership from the Instructional Science and Technology Community.

The second professional grouping, instructional technologists and those in training in human development who use similar methods, have a greater likelihood of providing leadership in assessment integrated with instruction than does the testing industry. Like the measurement professionals, not many of them have kept pace with developments in technology and few are strong in educational measurement. The new breed of cognitive and instructional scientists offer many exceptions to this generalization, at least as regards technology. The subgroup who have engaged technological problems most resolutely are generally known in the Integrated Learning Systems industry.

Current Status of the ILS Industry. An industry has been created for Integrated Learning Systems. It has grown out of earlier work in computer managed instruction and computer assisted instruction, including drill and practice systems. The industry is made up of a group of small specialized ILS companies and a group of large computer manufacturers. The small companies sell software, hardware, and services. They sell hardware as dealers, primarily, although one of them (Wicat Systems) began as a vertically integrated company that developed hardware, software, courseware, and provided services. Wicat Systems has since offered its educational courseware and services for delivery on commonly available PC platforms. The ILS industry is different from the industry for educational software that will run on stand-alone PCs. Stand-alone operation is a different concept from a system which integrates instruction, management, and testing through one networked configuration. Several of the ILS companies are systems integrators and developers.

[55] Winton Manning, Development of cloze-elide tests of English as a second language. ETS Research Report RR-87-18, Princeton, N.J: Educational Testing Service, 1987.

[56] Michael E. Martinez, John J. Ferris, William Kraft, Winton H. Manning, Automatic scoring of paper-and-pencil figural responses, ETS Research Report RR--90, 1990.

Michael E. Martinez, A comparison of multiple-choice and constructed figural response items. Paper presented at the meeting of the American Educational Research Association. Boston, MA, April, 1990.

integrate hardware, software, and service, and install the ILS systems in the schools as a part of their overall fee. This group of small specialized ILS companies includes Jostens Learning Corporation, Computer Curriculum Corporation, Wicat Systems, Plato Learning Centers, Wasatch Systems, New Century, and others. The distinguishing feature of this industry is the possession of substantial curriculum materials that operate on the integrated learning systems and cover entire subject matters over several years of the K through 12 curriculum. The set of small specialized ILS companies have found a market niche that is growing rapidly. Many labs and learning centers are being installed in schools. The group of small ILS companies probably accounts for about $300,000,000 dollars per year in volume, while the large computer manufacturers, of which IBM, Apple, and Tandy are the most prominent, control over $500,000,000. IBM is the most vertically integrated and largest of these and offers courseware, software, and hardware under its own label. Much of the courseware and software has been purchased or contracted from small software companies, from consortia, and from individuals.

The large computer manufacturers install labs consisting of personal computers of their own make. Sometimes these PC's are stand alone, but increasingly they are networked into a central file server for use in a variety of educational activities using computers. This includes instruction in computer science and programming, instruction in word processing, spreadsheets, and other business productivity software tools, writing labs integrated with writing instruction in a variety of classes, desktop publishing and graphics labs.

The Previous Evolution of Integrated Learning Systems. It was difficult enough to learn how to develop interactive instruction, then integrate it with CMI-like testing and with instructional management. Now forces in the marketplace, and voices in the scientific community, are calling for a better and deeper kind of assessment than the computerized item tests provide, and for a deeper form of integration with instruction.

Some of the evolutionary threads that have led up to today's ILS systems have a continuing influence today. The Plato system was developed in the College of Engineering at the University of Illinois, starting in 1969, under the direction of Dr. Donald Bitzer. Plato introduced some of the first multi-terminal labs at learning centers in schools. These were supported by large mainframe computers made by Control Data Corporation (CDC), but have in recent years been replaced by networked personal computers in the numerous schools and colleges using this system. There was a brief merger between Plato and Wicat to form the Plato/Wicat Company, and not too long after this merger failed, the Plato labs were acquired by The Roach Organization who markets and supports them today. Plato was championed for many years by William Norris, chairman and founder of CDC, who as head of the Norris Institute today is an statesman and leader in the movement toward transformed schools through the use of integrated learning system technology.

The Stanford Institute for Mathematical Studies in the Social Sciences pioneered drill and practice labs that originally used noisy teletypes (this worked extremely well in schools for the deaf). This work was soundly based and thoroughly researched, and led to the founding of Computer Curriculum Corporation, an integrated learning system provider that continues as a strong player today. The Stanford Institute also influenced the design of the IBM 1500 system, which in many ways was the earliest prototype of the twenty- to thirty-terminal integrated learning system labs found today. This system was discontinued in the early 1970's, but was very influential in building the human infrastructure for Integrated Learning Systems.[57]

The IBM 1500 system was one of the parents of the TICCIT system, completed in 1975 under National Science Foundation funding. TICCIT offered a thirty-two terminal integrated learning system. The other "parent" was the technology of interactive two-way television over cable, developed by the non-profit MITRE Corp. The TICCIT system was designed by the team that balanced the contributions of

[57] C. Victor Bunderson and Gerald W. Faust, "Programmed and Computer-Assisted Instruction", in N. L. Gagne (Ed.), THE PSYCHOLOGY OF TEACHING METHODS: The Seventy-fifth Yearbook of the National Society for the Study of Education, Part I, (Chicago, IL: University of Chicago Press, 1976). pp. 44-90.

instructional scientists/ technologists and engineers. It was built around a coherent instructional model and a coherent instructional management model. Hazeltine Corporation acquired TICCIT from the prime contractor, Mitre Corporation, after the NSF funding ceased. Hazeltine in turn sold it to Ford Aerospace. It was recently acquired by a larger training and simulation company. The TICCIT system used testing intimately integrated with instruction and with management, and had an early form of computerized mastery testing.

Another early ILS company besides Plato and CCC was Wicat Systems. The Wicat Integrated Learning System drew on some of the experience with the Stanford drill and practice systems and on the TICCIT computer-aided instruction system. Wicat systems also benefitted from some of the earliest interactive videodisc work. Wicat's current business consists of two parts: training systems, especially simulators for airline and other industrial training; and education systems, integrated learning systems installed in hundreds of schools. Wicat invested substantially in computerized testing and has several batteries of computerized achievement tests that may be integrated with most curricula. With help from foundation grants, Wicat's nonprofit institute also developed a battery of "learner profile" tests, which included many innovative item and test formats.

Labs for general computer use in schools represent another evolutionary trend. These labs have primarily been equipped with stand-alone computers used for programming, computer literacy, and now word processing, and other business productivity tools. So long as these labs consisted of stand-alone computers, they had little potential for computerized assessment or integration of assessment with instruction. However, as many of them have become connected to file servers with network software, it becomes feasible to use them as learning/assessment centers.

The Possible Future Evolution of Integrated Learning and Assessment Systems Figure 2 depicts some possible trends in the evolution of integrated learning and assessment systems. As mentioned above, stand-alone computers could not integrate management or measurement, but could give individual lessons and provide opportunities for tool use. The networked labs were of two kinds: integrated learning systems and specialized computer labs. Off to the side are special simulators, such as those found in driving classes and vocational classes, and two-dimensional simulation software programmed on interactive videodisc systems.

Currently there is a movement to place integrated learning systems in classrooms instead of in separate lab rooms. There are three forces driving this:

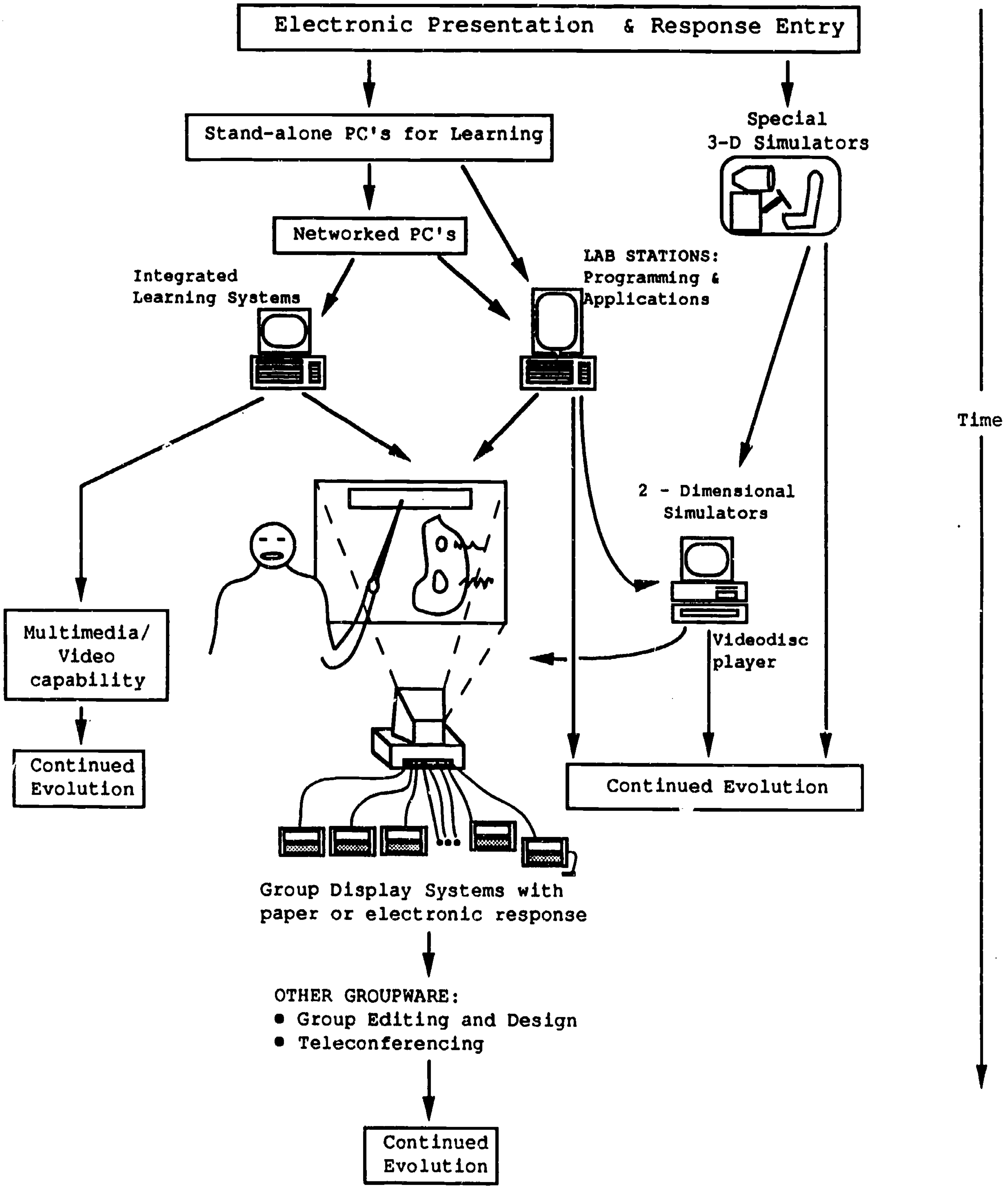
1) Educators have difficulty taking an entire classroom out of service (actually the service intensifies but this is not their perception) so they wish to spread the terminals out among the classrooms.
2) Educators feel that the teachers will become more involved if the computers are in the classroom.
3) Computer manufacturers can sell more hardware this way. For example, one thirty-terminal lab versus fifteen classrooms equipped with four terminals each presents an obvious short-term business payoff.

The thirty-terminal lab has many advantages for computerized educational assessment because of the security, standardization, and monitoring required in high-stakes testing sessions. The system administrator in the learning center can manage testing sessions. The group presentation computer at the front of the room has advantages for integrating instruction with assessment, and so does the learning lab. The most difficult configuration is classroom PC's. Four terminals in a classroom will go unused most of the time because the entire group must be attentive to the group activities. Texas Learning Technology Group has monitored the use of both teacher-controlled presentation systems and four or five classroom PCs. The group system achieves 45% utillization while the stand-alone PC's are used under 10% of the time.

Putting integrated learning system hardware and software as presently designed into classrooms is probably not a good idea. It carries with it a management model for individualized instruction, but classrooms are managed primarily for group instruction. Therefore, Figure 2 and Figure 1 both predict the development of group-paced assessment and instruction technologies. The teacher controls the computer with a file of visual materials, including videodisc simulations and scientific visualizations. Assessment questions are integrated with the presentations and students may respond either on answer sheets used for practice and feedback, or in the future, on small electronic devices like notebook

Figure 2

EVOLUTION OF COMPUTER - ADMINISTERED LEARNING AND MEASUREMENT SYSTEMS


Electronic Presentation & Response Entry
Stand-alone PC's for Learning
Special 3-D Simulators
Networked PC's
LAB STATIONS: Programming & Applications
Integrated Learning Systems
Time
2 - Dimensional Simulators
Videodisc player
Multimedia/ Video capability
Continued Evolution
Continued Evolution
Group Display Systems with paper or electronic response
OTHER GROUPWARE:
● Group Editing and Design
● Teleconferencing
Continued Evolution

computers. These devices have potential uses in individualized instruction, individualized testing, homework, and tool use, as well as group response.

SUMMARY

Current uses of computers in assessment have primarily been substitutive and incremental, but there are several transformational possibilities. This section has considered computer-administered item tests and incremental improvements to them, standardized performance tasks, computers as tools to get at process assessment and feedback, computer-enhanced display and interaction, and computers that integrate assessment with instruction and management. To achieve the benefits of any of these applications, whether incremental or transformational, a new infrastructure must be established in the schools to put the tools into the hands of the students and the teachers. Discussion of the probable evolution of both answer sheet systems and interactive workstations introduces us to the future of CEA, where group-oriented systems integrated with labs and with portables can play a transformational role. In Section III, we turn to further elaboration of these future possibilities.

SECTION III: FUTURE USES OF COMPUTERS IN ASSESSMENT

The last section showed current uses of computers in assessment are primarily substitutive and incremental in terms of the technology diffusion model. Transformational uses were introduced in Section II; these include the use of computers to administer standardized performance tasks, and to provide feedback and helps related to process during tool use. Other transformational possibilities included graphics, animation, and multi-media for deep visualization of difficult concepts, and perhaps most importantly, the integration of assessment with instruction and with management. This section will concentrate on the transformational applications and will discuss scientific and technological developments underway and anticipated that can lead toward these transformational possibilities.

TWO FUTURE SCENARIOS FOR COMPUTERIZED EDUCATIONAL ASSESSMENT:

Scenario 1. Three students are clustered around a video screen. Each holds a notebook-sized computer with an invisible infra-red data linkage to a computer that is displaying interactive graphics and video on the screen. They are managing a simulated McDonald's restaurant franchise in their city. Complex management issues are presented to them, with the emphasis on financial decisions. The students perform some calculations on their notebook computers, then signal their decision. Two agree, but one has a different decision. The system feeds this information back to the students, clearly displaying the differences in both decision process and result. The students argue for awhile, then agree with the one student and continue with the simulation. The two saw their different errors. The system, meanwhile, has recorded the arithmetic errors made by one of the students and the critical reading error that led another to use a wrong process. The system updated the learning progress map of the one in the domain of the mathematical concepts of fractions; and the critical reading map of the other. When each student re-enters their respective map a comment will be provided to each, designed to motivate practice and prescribe tutorial help with fraction and critical reading concepts, and advice on how to make better judgments in the future.

Although Scenario 1 can be seen as an integration-level test of rapid judgement requiring financial calculations, the students do not regard it as a test. No grading has occurred, but the experience is perceived by the students as an opportunity to debug their thinking skills. They were deeply engaged and motivated by practicing in a simulated area they had chosen. Traditional midterm and final tests are scheduled later in the semester, and the students know these tests will involve writing a report about the experience, working another unencountered problem alone in a simulated environment, and a 30-minute computerized mastery test of scaffolding level objectives. They knew they could practice any of these tasks in advance of the high stakes final exam.

One of the students, as it happened, was a high school junior working on a more extensive written report on the franchise simulation to put in her portfolio, as part of the admission requirements to a prestigious business school. Using the word-processor in her notebook computer, she added a few thoughts to her report, gleaned from this day's discussion with fellow students.

Scenario 2. A group of sixth grade geography students in the midwest are typing short responses to a series of questions presented on a video projector at the front of the room. The video involves NASA satellite shots. Answers are entered into their notebook computers via an infrared link to the main computer at the front of the room. Students know that their responses (along with those of a nationally representative sample of students at 499 other schools nationwide), will be used to evaluate the nation's progress in geography instruction.

Although the questions are short, they are clustered together into small sets with a meaningful context. The sets differ in cognitive demand and can be placed on a learning progress growth scale. The students are motivated to perform well because they know that tomorrow the teacher will review their results, will show the class' standing on national and state scales (updated overnight), and will go over common errors, neatly listed and sorted by the computer overnight. Those with the greatest desire to learn rapidly from their mistakes will go over the missed items (stored in their notebook computers) that evening at home.

The parents know that the students will be bringing assignments home recorded in their portable notebook-sized computer. The PTA has been working with the parents to encourage them to work with these students at home. The computers at school remain on after hours. Telephone and modem linkages enable parents to dial up for further information about the student's progress on learning progress maps that visually display the curriculum for the student that year.

Administrators know that they have to strictly avoid the use of learning progress data while holding the teachers or students accountable or grading them. They are content to know that an outstanding accountability test will be given at the end of each block, and one that will be reported to the state at the end of the year. They are comfortable in knowing that the curriculum is aligned with this assessment, and that the teachers can tell them how the progress data is looking. They have statistical machinery that allows them to predict how well each class will be doing as a group on the assessment at the end of the year. Moreover, they have formative evaluation information that gives them early warning that the classes might not be doing well on the annual assessment.

Curriculum developers use the national item analysis results gathered from 500 selected schools to update tutorial exercise sets in geography that provide hints and helps for each of the common errors found in the analysis of the over 10,000 student response vectors. These sets will go out to participating schools in both printed and electronic form, depending on the delivery system at a particular school.

Both scenarios are consistent with the recommendations in this report. To move successfully in this direction will require progress in scientific foundations, technological tools, and policy.

SCIENTIFIC TRENDS

Trends in Cognitive and Instructional Science

The first scenario described above incorporates some guesses about where Cognitive and Instructional Science will position assessment and instruction in the future. Progress is being made in the analysis of cognitive processes for use in instructional design, individual diagnosis, and feedback. It takes years of deep cognitive analysis in specific task domains (e.g., the franchise management task) to be sensitive enough to provide the diagnosis of the fraction error and the critical reading error described in the scenario. The use of mass response collection methods illustrated in scenario 2 might lead to sets of common errors that will permit a more pragmatic, if less precise, approach to the problem of building a responsive and reasonably intelligent learning environment with built-in diagnoses of the most important pitfalls.

The power of "impasses" -- the evidence presented in a problem situation indicating a selected solution approach does not work -- is understood in cognitive science. Impasses are opportunities to learn. Students are challenged to try using a better approach. If the system contains help so powerful that all with desire can learn, at least the most common impasses will be followed by hints, cross-references, or examples to guide thinking toward that better way. Current intelligent tutors that can generate the best help for each learner are even more utopian than what is envisioned here. We predict the development of empirical methods that can aid in discovering the important impasses quickly, then lead to tutorials that teach one or two correct approaches. We do not envision a "super-diagnoser" of any possible error model that a confused but creative mind may have generated.

The power of social context and group effects is very much on the scientific agenda for cognitive science today, and scenario 1 honors this by using group discussions among the three students, as well as on-to-one interactions. Meaning that will transfer to new situations is "negotiated" through group dialogues (and inside one's own head, as in the reflection afterward by the girl working on her portfolio). Useful knowledge structures are not memorized from books or computer screens.

Respect for integration, transfer, and creative production also characterizes cognitive science today. In scenario 2 even the scaffolding items of geography instruction were integrated into *testlets* containing more meaning and context than individual items. In scenario 1 a standardized simulation task was used to teach an integration objective. The impending final exam question using an unencountered problem would provide a test of knowledge transfer. The portfolio building activity described for one stu lent offered creative production and transfer.

Learning and Cognition are situated in real human and physical contexts[58]. Providing more project-like activities, and small working groups introduces more of this reality into school settings.

Metacognition and learning strategies are always difficult to address. The situation is improved by introducing simulation exercises and computer tools. Learning strategy is a dull subject without a specific context, without specific "rules of some game". Generally speaking, about the best one can do is some variant of "SQ3R" (Study, Question, Read, Recite, Review). Strategies become vital and interesting within the context of some "game" like chess, franchise management, using an outline processor or a spreadsheet program. Interdisciplinary expertise can be shared, and strategies and tactics for accomplishing different goals can be discussed within these specific domains. The very discussion thus generated is metacognitive -- reflecting on thought. Specific tools and tasks with a formal structure and syntax can be used to promote this kind of discussion and self-reflection. That is why the proposed emphasis on strategy objectives, practiced and perhaps eventually assessed within performance tasks and computer tool use is recommended.

Conative and self-management objectives are at least as important as cognitive objectives. Conative objectives deal with motivation, persistence, commitment. There must be research and development aimed at breaking out of the cycle of dislike and avoidance for conventional testing and instructional practices. Students need help and confidence-building experiences leading up to challenges that they respect and value. The two scenarios depicted the achievement of new attitudes toward finding ones own errors -then fixing them. New student attitudes toward self- management and self-motivated writing, reflection, and learning were also depicted in both scenarios. The reasons for claiming that CEA can help are given in the section on achievement constructs, below.

Trends in Measurement Science

In scenario 2, two recent developments in measurement science are assumed. One is a national scale of learning progress. A candidate for developing such scales is the Hierarchically Overlapped Skills Test (HOST) developed by Don Rock[59]. It is based on clusters of items (sometimes called testlets) that contain increasing levels of cognitive demand, each one including the previous level. The constructs underlying each level are understood by the testlet developers and can be illustrated and taught. A student's position on the scale, unlike a norm-referenced ranking scale, has instructional utility. The student knows what he/she can do, and what must be learned next. Other new scales diagnostic of learning progress are also under development[60].

The theory and practice of using testlets (instead of individual items, which do not carry enough context) is the second recent development[61]. Measurement science has developed powerful machinery

[58] John Seely Brown, "Toward a new epistomology for learning", in C. Frasson and J gauthiar (eds.), Intelligent Tutoring Systems at the Crossroad of AI and Education. Norwood NJ: Ablax, 1989.

[59] Rock (1989)

[60] Yamamoto and Gitomer, 1989.

[61] Howard Wainer & G. L. Kiely, "Item clusters and computerized adaptive testing: A case for testlets". Journal of Educational Measurement, 1987, 24(3), pp. 195-201.

Howard Wainer and Charles Lewis, "Toward a psychometrics for testlets", Journal of Educational Measurement, 19--

for dealing with item tests. However, it has not yet developed equivalent machinery for measuring constructs inferred from the highly variable responses made during a flexible performance task, like a computer-controlled simulation, nor during tool use while producing an exhibit. Item clusters may use familiar item types, yet group them into contexts that can represent levels of cognitive demand, subsets of content or process, aspects of procedure, aspects of strategy. Banks of such clusters can be sampled using adaptive techniques to gain much valuable knowledge about how each student is progressing.

Developing models for incorporating time as a measurement variable constitutes one of the most important trends in measurement science. Many aspects of time may be measured, including the time it takes a student to complete a particular task, the time it takes to complete a whole set of tasks, the amount of time something is displayed for a student to observe. In the past, descriptive statistics on time intervals have not been kept because of difficulties in precisely timing individual items and component parts of items. Today, with computerized testing and computerized administration, measuring time is easy. However, little research has been done in this area. The relation of time to all sorts of proficiencies in many fields and for many thought processes is a subject for research that will occupy thousands of graduate students and scientists in the future.

Research Integrating Assessment and Instruction

Successfully integrating assessment with instruction has been the goal of measurement scientists for decades. Enormously difficult to accomplish, it involves the integration of many scientific disciplines: Developmental Psychology, Cognitive and Instructional Science, human and organizational disciplines like Organizational Behavior and Anthropology, as well as Measurement Science. Since science and measurement are inseparable, it is vital that the interdisciplinary nature of the field be recognized and actively promoted by today's educational measurement science leadership.

In the past six years, there has been a flurry of activity in the field of educational measurement around the integration of assessment with instruction. For example, within Educational Testing Service in Princeton, N.J. there is a recognition that this integration represents a fertile area for research. Benefitting from a broad range of consulting expertise and collaboration with universities and private organizations, some promising new developments have emerged. A number of these are described in two forthcoming books[62]. In addition, A new monograph[63] commissioned by ETS proposes a research and development agenda for integrating assessment with instruction. Through an extensive literature review and analysis, Richard Snow of Stanford University and Ellen Mandinach of ETS seek to identify what can now be answered through research, organized around four general questions:

1) What constitutes learning progress toward mastery in an instructional domain?
2) What constitutes diagnostic assessment of learning progress for instructional use?
3) How might performance tasks that provide such assessment be designed and evaluated?
4) How might collections of performance tasks be mapped into an instructional domain to guide instructional adaptation?

The recommendations Snow and Mandinach make for Systems That Integrate Instruction and

[62] Frederickson, N., Glazer, R., Lesgold, A., and Shafto, M., Eds. (1990), Diagnostic Monitoring of Skill and Knowledge Acquisition. Hillsdale, NJ: Erlbaum

Frederickson, N., Mislevy, R., and Bejar, I. (in press) Test Theory for a New Generation of Tests. Hillsdale, NJ: Erlbaum.

[63] Snow R. E., and Mandinach E. B., (1990, Pre-publication draft), Integrating Assessment and Instruction: A Research and Development Agenda.

Assessment (SIIA) are consistent with several of the major recommendations made in this paper. In particular, research and development practice must move as rapidly as possible in the direction of learning progress help systems. Snow and Mandinach further point out that in the old paradigms (e.g., CMI) the instructional tasks and tests remain distinct from instruction. They are hopeful that the field is maturing to the point where integration of the two can be successfully addressed. However, they point out with needed caution that theories of learning progress and its diagnosis for teaching in an instructional domain are not now available. Their agenda for research and development is therefore a long-range one.

Measurement science will be challenged as never before to stretch into previously uncharted research areas. New kinds of growth scales are needed based on sophisticated cognitive and developmental theories, rather than the overly simplistic and erroneous model that curriculum is comprised of equivalent snippets of knowledge. Moreover, scoring methods for complex integrative tasks, performances, and tactical sequences in tool use will open up new possibilities for computerized assessment. Promising models that mix growth measurement with diagnosis of different strategies are now being developed. In a way, measurement science is entering an era pregnant with possibilities for the explosion of new models and methods, much like the era of L.L.Thurstone when many new methods for the testing of primary mental abilities were pioneered.

FUTURE PROGRESS IN INSTRUCTIONAL SCIENCE AND MEASUREMENT SCIENCE DEPENDS ON THE DEVELOPMENT OF GENERALIZED ACHIEVEMENT CONSTRUCTS

Educational and psychological tests measure human aptitudes, traits, or achievements. None of these are visible physical quantities that can be observed with the five senses; therefore, before appropriate items or performance tasks can be developed, we must "construct" a mental picture with words or images to define what we are trying to measure. These ideas are called "constructs." In aptitude testing, psychologists and measurement scientists have developed such constructs as spatial ability, induction, and deductive reasoning. Psychological tests measure constructs of personality and of clinical pathology. Admissions tests measure verbal and mathematical ability developed over previous learning and schooling. In achievement measurement, there is no generally agreed-upon set of constructs, yet differing conceptions have a great impact on research, teaching, and on how the public views educational goals[64].

As was mentioned earlier, the predominant uses of educational measurement earlier in this century were for sorting and selection based on intelligence or aptitude. The historical shift from sorting and selecting to promoting growth in learning is correlated with the shift away from aptitude measurement and toward achievement measurement. It is worthwhile to understand how aptitude constructs and test formats influences achievement testing today.

Contrasting Aptitude Measurement and Achievement Measurement

Aptitude measurement cuts across many contexts. When we are measuring an aptitude or developed ability[65], it is appropriate to sample from many small performances with little specific context. Decontextualized items get at abilities that are general -- they are applicable to many contexts. Aptitude testing is certainly one of the most successful and widespread applications of behavioral science, at least

[64] Nancy S. Cole, "Conceptions of educational achievement", Educational Researcher, April, 1990, Vol. 19, No. 3, pp. 2-7.

[65] College admissions tests, of which the Scholastic Aptitude Test (SAT) is the most widely used, are no longer said by their makers to measure aptitudes -- innate and difficult-to-change traits of individuals, but "developed abilities." A "developed ability" is subject to schooling. It is assumed that the SAT measures verbal and mathematical abilities that have been gained through studies and exercises in many subjects, as well as by activities outside of the school, taken over the twelve years of schooling. Developed abilities take many months and years to develop.

when success is measured by how widely it is used in military, industrial, and educational settings. The widespread use of aptitude testing has generated much criticism about its validity, equity, and inappropriate uses. From the perspective of this paper, one of the most damaging impacts of aptitude testing is that its very success has promoted extensive copying of its item types and methods in the quite different field of achievement testing.

Aptitude testing methods should not be copied for tests of achievement. The tendency on the part of teachers (taught to them in college measurement courses) to copy the item types and test formats from aptitude testing for use in classroom achievement testing. Aptitude test formats carry with them as hidden conceptual baggage the "high-stakes" frame of reference (ranking and grading) instead of the help service mentality. It also results in a definition of the curriculum as actually delivered to the students that is more fragmented and decontextualized, not integrated and clearly relevant to valued total performance.

Achievement Measurement is Always in Context. Unlike aptitude constructs, which utilize decontextualized items, achievement measures are always in a subject-matter context and a value context -- valued performances worth doing. We achieve at valued performance tasks within the domains of history, science, art, automotive maintenance, etc. The public gives "face validity" to curriculum tasks that resemble valued tasks or products in life and work, but the prevailing theory in educational circles is that we should not make the context too specific, else we are "training" rather than "educating". Leading cognitive scientists are now insisting that valuable and persistent learning is always situated in a meaningful context.

Achievement Measures are Derived from a Curriculum Plan, and from a Set of Values. Theories and values about what knowledge is lead to definitions of an achievement domain, and a curriculum is a map of what must be covered, and in what rough order, within that domain. Typical practice when specifying a curriculum is to begin with content outlines (topics, not performance tasks), then develop a set of objectives. These reflect the prescriptions for good performance objectives found in instructional technology text books. Unfortunately, these objectives tend to emphasize too strongly simple verbal knowledge rather than integrated performances. Later, when teachers are confronted with a fairly extensive curriculum outline, usually provided by their district, they pick out the objectives they feel they can cover, snipping them, as it were, from the horizontal fabric of the "quilt" that represents the topical domain. At worst, this practice results in a "snippet curriculum" where little pieces of factual knowledge are presented unconnected to one another and unrelated to the complex and nested thinking processes used by those knowledgeable in the domain. This is, of course, a caricature of the worst in curriculum implementation. The curriculum outline built by this process, may be quite thoughtful and integrated (but usually it has far too many topics in it). As implemented by busy teachers who must read the topical outlines and make selections to fit into the time available (and to fit what they are knowledgeable to teach), the best curriculum guide may be reduced to a "snippet curriculum" when implemented in a particular class.

Even the best curriculum guide that follows the practice of writing performance objectives of a primarily verbal nature may produce tests and instruction that are far from what is desired and needed by our culture of the 1990's.

Achievement Constructs are Needed in Order to Develop Better Instruction and Measurement Instruments. Instructional scientists have been groping toward a generally agreed-upon set of constructs that cut across subject matters and permit the prescription of the more promising approaches to teaching and learning. Taxonomies of educational objectives have been the most visible manifestation of this search for constructs useful in instructional development.

Curriculum developers have been guided by two main taxonomies of objectives. An objective is a prescription for writing performance items or tasks, so the objectives selected, and the constructs behind them, determine the measurement instruments used to assess achievement in a curriculum. One such taxonomy of educational objectives was developed by Benjamin Bloom and his co-workers[66]. It includes such constructs as memory, application, analysis, and synthesis. Another taxonomy is based on Gagne's[67] model which includes constructs such as memorization, concepts (classification behavior), rule using, and problem solving. Instructional developers who use these taxonomies are likely to develop objectives at higher levels of processing than those who use no model or taxonomy, only prescriptions for writing behavioral objectives. These prescriptions have an implicit model of achievement in them, and this implicit model is that knowing verbal information "about something" is the predominant objective or kind of behavior that can and should be measured.

The problem for all developers is that there is a paucity of examples of test item and task types that can be developed and used within the constraints of the print delivery system. The item types in the developers workshop of tools are inherently dominated by short instructions on a printed page, and by short quickly recorded and interpreted responses on another printed page. Aptitude test items offer a strong example for developers to follow.

CEA is Fundamental in Measuring and Teaching Achievement Constructs. Multimedia presentation and dynamic interaction capabilities of Computer Administered Tests are more fundamental in getting at different achievement constructs than would be seen at first glance. They are far more than a more interesting and motivational presentation.

This paper adopts the perspective that the measurement methods of standardized performance tasks, student-developed exhibits, and process measures during performance permit curriculum developers to tap processing well above the level of the typical print-mediated verbal test item. *This is the crux of the argument for using Computerized Educational Assessment: CEA takes us beyond the print delivery system to scorable interactions with more realistic, contextuallized tasks. Such tasks are necessary in assessment of more important, higher-order constructs than can be measured and implemented using the print delivery system alone.*

Table 8 is the embodiment of this central argument. It adds another dimension to the four measurement methods: the dimension of generalized constructs to describe individual achievement. We will call these "achievement constructs" and note that they are related to curriculum objectives at a level of generality higher than is found in the familiar taxonomies of educational objectives.

[66] Benjamin Bloom, Taxonomy of Educational Objectives. Handbook I: Cognitive domain, (New York, NY: McKay, 1956).

[67] Robert Gagne, The Conditions of Learning, 4th ed. (New York, NY: Holt, Rinehart and Winston, 1985).

Table 8
Measurement Methods
for Different Achievement Constructs

ACHIEVEMENT CONSTRUCTS (Kinds of Objectives)	MEASUREMENT PRACTICE			
	Item Test	Performance Tasks	Student Exhibits	Process Measures
Scaffolding Knowledge	****	**	*	*
Integrated Performance	*	****	**	***
Creative Production (Transfer of Integrated Skills)	*	**	***	***
Strategy Improvement	* to (-)	***	**	****

The four achievement constructs found in the four rows of Table 8 are defined as follows:

1) **Scaffolding knowledge** is verbal knowledge *about* some topic (terminology, definitions, classifications, rules).

2) **Integrated performance** is the capability to perform well in more complex tasks requiring the student to use or apply the scaffolding knowledge, along with common-sense and task-specific skills, to solve some problem, operate some real or simulated equipment, perform some experiment, write up a carefully specified paper, or make a carefully specified presentation. Standardized performance tasks generally take far longer to complete than short verbal items.

3) **Creative production and transfer of performance capability to a new situation.** Creative production objectives require students to transfer creatively both scaffolding knowledge learned previously, and previously demonstrated ability to integrate into a new situation. The project task requires them to use their knowledge and skill to design and produce some written or mediated presentation or some product.

4) **Strategy improvement.** Strategies are of two kinds: learning strategies and performance strategies. The former applies to techniques students use when confronted with learning tasks for the different types of scaffolding knowledge, or learning to perform an integration task well, or learning to create a new kind of exhibit. Performance strategies are specific to some well-defined task, game or tool to be used. Strategy improvement is a construct that must be inferred from more proficient learning of a particular type, and from more efficient choice of strategy and tactics in process of performing or producing.

All four achievement constructs have both cognitive and conative aspects. The cognitive aspects include knowledge and proficiency. The conative aspects include motivation, commitment, desire, and persistence for achieving at high levels. Conative objectives are of utmost importance. We often pay lip service to them but do not try to assess them. The importance of conative objectives leads to a corollary to the central argument of this paper: *Instruction with integrated, unobtrusive measurement using all of* *the four kinds of performance tasks will be more motivating and less aversive to students, offering promising new approaches to the achievement of conative objectives.* Consider the negative attitudes and avoidance toward testing in particular and toward studying in general (found among many of today's youth). Can a new family of help systems with interesting, challenging performance tasks and assignments to produce exhibits lead to more engaged and challenged students? As another possible benefit, can unobtrusive measurement (e.g., measuring use and avoidance) help identify attitudes toward learning and persistence, leading to ways to improve them?

As with the achievement objectives, the use of computers in achieving conative objectives is fundamental, not incidental. There are several reasons for this. First, it has been demonstrated repeatedly that interactive instruction, especially where video discs and video are involved, is very engaging, interesting and motivating to most students.[68] Second, the miss-placed emphasis on "getting the right answer the first time" (a no-win situation) might change. Most students regard an error as a sign of inadequacy, rather than as an immediate opportunity for unqualified and rapid self-development. Some educators have similar defensive attitudes. While teachers still have to create an environment to permit a different conception toward errors, computer use presents a different model. Finding a bug or error in a computer program is a part of the process, not a reflection on the individual. To find a bug is a welcome thing; perhaps it means that you have now caught that last trash fish out of the deep trout pond. Perhaps the debugging metaphor can improve attitudes towards learning from errors. Third, the use of computer productivity tools to develop student exhibits brings with it direct instruction about the need for revision. Once again, the students are taught actively that errors are not bad, but a part of the process. Finally, good performance in complex tasks, and products that can be exhibited in a portfolio bring validation from adults and peers, not just the questionable validation of a grade.

TECHNOLOGICAL TRENDS

Predictions of Technology Trends

The literature on future technology trends is very large and spans both popular magazines on personal computer trends, and magazines and journals associated with computer science and engineering. Among the latter of the IEEE Spectrum (Institute for Electrical and Electronics Engineers) has an annual update on technology trends.[69] This large and varied technical literature cannot be reviewed here, but some themes clearly have potential significance for the field of assessment:

o Group process and team building activities will become more important. The increasing use of interactive computer technologies by groups is much talked about; including networking, electronic mail systems, and the like. "Groupware" is a new buzzword. Groupware is defined as:

> *Computer-based systems that support groups of people engaged in a common task (or goal) and that provide an interface to a shared environment.*[70]

It can refer to face-to-face interaction that occurs at the same time in the same place, or it can refer to interactions that occur at different times and in different places.

Education is a field well positioned to take advantage of groupware. It is organized around

[68] Gary Borich, The Texas Learning Technology Group Evaluation Report, Austin, Texas, 1988.

[69] "What's new in products and applications", IEEE Spectrum, The Institute for Electrical and Electronic Engineers, Jan. 1991. It contains news of a 64 million bit memory chip that will be in mass production for products such as notebook computers in 4 years. See also "Fifteenth anniversary summit", Byte, McGraw Hill Publications, September 1990. To commemorate its fifteenth anniversary, Byte magazine assembled computer industry leaders to make their forecasts of personal computing.

[70] Ellis, C.A., and Gibbs, S.J., and Rein, G.L. Groupware: Some Issues and Experiences. Communications of the ACM, Volume 34, #1, January 19, 1991, pp. 39-57.

classrooms; rituals and traditions are deeply ingrained, such as appearing in class at a scheduled time, and communicating verbally and nonverbally with other students and teachers. Effective groupware could be built around interactive display generators and projectors, with files of materials for explaining difficult concepts, using animations, scientific visualizations, games and assessment tasks that can be displayed and discussed by the class. Students equipped with handheld response units add yet another dimension to the potential group experience. We have developed Scenario 2 to incorporate this trend toward group interaction. The ramifications for group problem solving, team building and gaming are virtually endless when technology is employed in this manner, but as yet, it is an area that remains poorly developed in education.

o Use of computer networks will become more widespread, easy to use, and increasingly connected over long distances. This is promising for record-keeping and statistical functions in CEA Systems. Stand alone personal computers do not lend themselves to CEA. The expansion of networks opens up the possibility of access to assessments stored at numerous host computers with delivery to individuals in homes and workplaces. Because of the security and privacy issues associated with high-stakes assessment, the kinds of assessment made available in this manner will most likely be help systems where disclosure is not an issue.

o Multi-media information display and interaction is a booming trend now and will continue. Within two decades, the world has witnessed a remarkable development in graphics, color and video display capabilities. High resolution interactive graphics, video, animation, and graphical user interfaces are becoming familiar to millions of computer users. Visualization techniques offer a new window on the previously unseen[71]. Speech recognition is an important goal that will become extremely important for assessment and instruction. Digitized video on magnetic disks is a current reality that will become more available and inexpensive in the future. A merging of video and computer technologies is expected, as well as the integration of computer and communications technologies.

o Portable systems, especially notebook computers, will continue to decrease in size and cost, and increase in power. As discussed in connection with Figure 3, the notebook computer offers this same portability and versatility, and for more functionality than printed answer sheets. Notebook computers are sometimes equipped with handwriting detectors, minimizing the importance of keyboarding skills in computerized assessment.

o Computer tools will extend our definition of literacy. Widespread accessibility of low-cost personal computers and production software like word processors, graphics design tools, spread sheets, equation solvers, etc. will increase and will define a new kind of literacy. This new literacy will become a part of the definition of an educated citizen in the modern world.

ADMINISTRATION OF FUTURE CEA SYSTEMS.

Trends in Paper Testing.

Fine grained scanners and advances in image processing will open up a new range of scannable item types beyond multiple choice. Several of these were discussed in Section II, including figural response

[71] Eric L. Schwartz, "Computing the anatomy of the brain", Pixel, the Magazine of Scientific Visualization, Vol.1, No. 1, Jan/Feb 1990, pps 20-27. In the premier issue of a new magazine on this subject, visualized neuroscience models "takes us into the skull" to give us insight into the brain's map-making ability.

G. Comer Duncan, "Visualizing the collision of a star with a black hole", Pixel, the Magazine of Scientific Visualization, Vol. 1, No. 3, July/Aug 1990, pps 24-29. A short time later, another visualization takes us out into the distant universe to "see" a rare astronomical event.

items (marking, drawing lines and arrows) and Clcze-elide (marking key words or inappropriate words in a sentence or paragraph). These item types are easily implemented using interactive computers. Portable answer media, in particular scannable answer sheets, are by no means restricted to 4- or 5-alternative multiple choice items.

The following response formats include those categories of response entry which can be accomplished on both answer sheets and electronic media, and those only possible on electronic media.

Response entry

Pointing/selecting responses:

o Select from a limited set of choices by darkening a box on an answer sheet, or point to one of several choices and click with ouse, cursor, moving field, or finger. Type the number or letter associated with the alternative selected on a keyset.

o Mark a part of a picture or line out a word on a high-resolution scannable answer sheet (Also possible on a bubble answer sh .t if objects and words are carefully positioned over the bubbles). Use a mouse, joystick, or trackball to do the same things on a computer controlled display.

Enter character strings:

o Grid-in a few digits and symbols on an answer sheet, or type the same on a keyset.
o Print by hand on high resolution answer sheets, or print with a stylus on a notebook computer or sensitive pad.
o Use a keyboard to enter character strings.
o Move a cursor among a set of letters and words, selecting them and dropping them to a line where the response string of characters is being build. Signal when completed.

Draw lines or shapes:
o Recognize line segments or directional arrows on high-resolution answer sheets.
o Recognize line and shape information drawn with mouse, light pen, finger, or stylus on a sensitive surface.

Vocal utterances:

o Digitize and recognize a given set of words.

Duration of response:

o Measure and record the latency before the response begins (thinking time).
o Measure and record the duration of the response; its entire composition.

Measure pressure, velocity, and direction:

o Use joystick, steering wheel, or other special interface device to input force, velocity, and direction information into the computer, and adjust displays accordingly to provide immediate visual feedback to the user.

Trends in Interactive Electronic Devices

All the response possibilities discussed for high-resolution answer sheets, vocal utterances and response times are possible on Educational Testing Stations as well as Laboratory Workstations. In addition, the display options that follow are increasingly being made available. Expense is a major factor in both display and respon e options, but costs have been decreasing rapidly in the memory requirements to generate the displays, in display drivers and monitors, and in response devices.

Trends in Information Display

Text: Multiple character fonts are available and the future will see a huge array of standard and customized fonts. Students can reflect their own stylistic "signatures" in student products. Graphical user interfaces assure that "What You See Is What You Get" (WYSIWYG).

Line graphics: Increased resolution, drawing, and animation capability will put drawing power into the hands of students and teachers for illustrated exhibits. Color displays will become easier to produce, edit, and display .

Photographic still images: These are increasingly easy to scan, digitize and manipulate for student exhibits and teaching materials.

Digitized Audio will be increasingly inexpensive and available for musical backgrounds (and foregrounds), and voice reproduction. Synthesized voice and music will be so available and inexpensive that it can be used increasingly both by materials developers and by students for their own exhibits.

Video (motion): "The technologies of video and interactive computers will merge[72].

Control of time intervals in displays will be used for many instructional relevant purposes, e.g.: testing of aptitudes for perceptual speed; pacing for practice in tasks in which speed is valued (e.g., typing).

Richness and realism in performance tasks and simulations will increase greatly in response to the dramatically improving capabilities of user interfaces. These capabilities, when coupled with the low cost and wide availability of "desk-top multimedia publishing", will enable more educational groups to develop their own materials, or at least to customize existing materials. Indeed, individuals "in their garages", having knowledge and experience in some valuable area, (perhaps acquired as a hobbyist), will be able to develop performance tasks and games with potential for instructional use. This will occur independently of the authors having knowledge or expertise in cognitive, instructional, or measurement science. Thus the results will be of mixed quality.

Notebook computers for each student

The trend toward reduced size and cost and increased capability has frequently been noted in the press, with innovations appearing almost monthly in computer magazines and newspaper ads. Computer companies will continue to increase notebook capabilities so that the price will not have to drop too low. For educational applications, modest processing capability and modest resolution monochrome displays can be produced relatively inexpensively.

This will make it possible to provide students in the classroom with response devices that can be used in a variety of ways:

1) To answer questions (i.e., Keyway system) on a printed worksheet, at the student's own pace. The notebook can then be taken to a file server at the front of the room for immediate scoring.

[72] Doug Engelbart, BYTE, Sept., 1990.

2) Utilizing telecommunications technology or a computer network, group assessment questions can be projected onto a large screen at the front of the room by the teacher, who can access an archive of interesting display and assessment tasks. As students enter in short answers or selections to limited choice items, limited choice or figural response items, the statistical results can be displayed to the whole class. When the items are a constructed response, the teacher can highlight and project particular answers that are worthy of group discussion.

3) As in Scenario 2, these group response devices can be used to collect pretest data, calibration data, and formative evaluation data from a sample of classrooms around the country.

4) More speculatively, with two-way response and a microphone, students could receive copies of visual materials, stored in a notebook computer along with a digitized version of the teacher's presentation. Each student could then go back and review the group activity using the hypertext graphical user interface, insert additional notes, edit existing notes, and erase what is not interesting.

5) Also more speculatively, at the end of the day, the teacher could transmit individual homework assignments from a teacher's workstation to individual notebook computers. These tasks could be customized to the needs of each student, through individualized learning progress calculations made by the computer.

The evolution of notebook computers for the four tasks listed is considered an important, but distant reality in the evolution of CEA systems. Groupware for education will continue to evolve, and will also include some of the features of special lab work stations, interactive video simulators, and even special simulators that could generate rich and highly effective instructional displays.

Special Simulators

As discussed in Section II, simulators are currently in heavy use in both military and industrial settings. They are highly cost-effective when a hierarchy of simulators is used to replace instructional time spent in actual equipment or in more expensive simulators. Thus a typical hierarchy of cost and complexity in pilot training for advanced aircraft is:

- At the top, it costs $5,000 per hour for flying the aircraft.
- Next, it costs $700 per hour for time in the four-dimensional movement simulator that moves about, tilts, accelerates, and simultaneously presents a visual display of the ground and sky while a trainee operates actual controls in the realistic cockpit.
- Third level position trainer costs $200 per hour. Here the pilot sits in a nonmoving mockup of the cockpit and learns the position and function of dials and controls.
- $20 per hour is the cost for time spent in the two-dimensional videodisc simulator. This is a personal computer controlling a color display and videodisc player. Many tasks at the scaffolding level and some integration tasks can be practiced in the environment of the two-dimensional simulator.

The cost benefit calculations with hierarchies of simulators is easy to compute. The more training objectives that can be accomplished in the lower cost simulators without loss of effectiveness, the better. It is now the case in the training of commercial airline pilots that their first flight in a large aircraft is with passengers. Their experience in the hierarchy of simulators has been shown through the experience of many trainees to prepare them adequately for their first flight as a copilot.

Note that validation of the objectives at each level is accomplished through performance tasks at the integration level in the next higher simulator. That is, one simulator is validated by time reduction in the next most complex simulator. The final step is the actual work itself.

It is interesting to speculate about what the future may bring for educational assessment. It is frequently the case that used hardware can be donated to schools by industry (only that which is not too costly to maintain). Vocational schools can be re-tooled to include and maintain costly manufacturing equipment, robotics equipment, etc. donated by industry. In a majority of classrooms, however, we will be fortunate to move up to the computer-controlled videodisc player to be used by the teacher in a group mode, and by a few students at appropriately scheduled times during the day and evening. Excellent simulations of complex integrative tasks can be administered on these devices. There might be a hierarchy of less expensive electronic devices below the simulator which prepare students for integration tasks presented on videodisc-based simulators.

Standardized performance tasks

In the discussion of trends in multimedia display and response capabilities described above, the opportunity for developing many more standardized performance tasks was discussed. It should be mentioned in this respect that advances in software methods will make this more and more plausible. Not only will "multi-media desktop publishing systems" become available to produce more standardized performance tasks, but advances in software development like object-oriented programming will help. Full performance tasks, or parts of them, could be programmed as reusable objects that could be used in different ways in the production of modified or improved performance tasks.

Unfortunately, as educational games and simulations are developed by those without adequate background in instructional design and measurement science, there will be a plethora of interactive performance tasks that do not fit well into any curriculum and that violate important principles of instruction and motivation, without clear instructional links. As Lepper and his colleagues[73] have pointed out:

"Their designs have often violated sound principles of learning and motivation theory. Without clear instructional links, they become discovery learning problems. Many students are unable to benefit from the implicit instruction."

Expert systems techniques offer considerable promise for future scoring of computer-based performance tasks, and for computer-generated diagnostic feedback during the process. A collaboration between measurement scientists at Educational Testing Service and Artificial Intelligence experts at Yale and Michigan has led to promising demonstrations in the field of computer programming. High school students learning the Pascal Programming language were assigned standardized tasks calling on them to write a program. An expert system program called Proust scored a set of such programs previously scored by humans in the Advanced Placement Test for Computer Science. The expert systems were able to produce scores for between 82% and 95% of the solutions, with high agreement with a human reader on the correctness of the solutions[74]. This team of researchers is seeking ways to integrate this powerful computerized assessment model with instruction[75]. Artificial intelligence methods are used to provide partial-credit scores and diagnostic analyses on each item the students work. Cognitively-based measurement models are used to generate diagnostic statements based on commonalities in performance

[73] M. R. Lepper, and T.W. Malone, "Intrinsic Motivation and Instructional Effectiveness in Computer-Based Education", in R.E. Snow and M.C. Farr, (Eds.) Aptitude Learning and Instruction, Volume 3, Conative and Effective Process Analysis, (Hillsdale, NJ: 1987). pp. 255-286.

See also E. B. Mandinach, The Role of Strategic Planning and Self-Regulation in Learning and Intellectual Computer Gain, Published doctoral dissertation, (Stanford, CA: Stanford University, 1984).

[74] Henry I. Braun, Randy Elliot Bennett, Douglas Frye, and Elliot Soloway, Scoring constructed responses using expert systems, Journal of Educational Measurement, Summer, 1990, Vol. 27, No. 2, pp 93-108.

[75] Randy Elliot Bennett, Toward intelligent assessment: and integration of constructed response testing, artificial intelligence, and model-based measurement, Research Report RR-90-5, Educational Testing Service, May, 1990.

across tasks. Models for diagnostic feedback back come out of work with expert systems[76]. Scoring of complex constructed responses to standardized tasks in mathematics is another area showing promise[77]

Tools for Student Products

In the OTA technology diffusion paradigm, the third stage is to go beyond substitution and incremental improvement to the introduction of whole new concepts. Process measures during tool use as part of a help system are perhaps the premiere example presented in this paper of such an evolution.

Students increasingly use outline processing software, word processing software, and desk-top publishing to produce their own creative products. The classroom environment is suitable for continuous measurement and feedback. Thus, virtually all of the students' responses and response times while using software tools can be monitored for the purpose of providing learning progress help. In this case, measurement can be used to monitor the use of different tactics and to suggest sequences of commands or activities that the students have overlooked, forgotten or not learned. Before on-line measurement is in place, the students can print out intermediate and final products and submit them to teachers and other graders, including fellow students, for constructive feedback. This peer review models the process that professionals often use in developing their own products in the workplace. Ratings and feedback can be abstracted by the students and summarized in their own portfolios.

In the emerging portfolio culture now being developed in leading schools, portfolios include not only favored completed products, but also personal journal information where the students learn to audit their own learning processes during production of exhibits. These reflections also become part of their growing portfolios.

Two examples of promising systems developed for high stakes testing, that nevertheless demonstrate the potential for integrating assessment with tool use are under development at Educational Testing Service. The first of these deals with the simulation test being developed with the National Council of Architects Review Board (NCARB), the second is a word processing test being developed by ETS in cooperation with the KEE corporation.

In the NCARB simulation, a computer-aided design-like system is provided to the candidate for an certificate in architecture. The system is very user-friendly and resembles the office of a practicing architect. A reference book can be selected from a shelf on the desktop to review statistics, construction standards, etc. On a separate screen, the architect can use some design tools to begin laying out a house design according to a standardized specification given in the test instructions. There is a lot of flexibility in the way any design can be approached, as long as it meets the overall specifications. This falls under the category of a standardized performance task, but it is implemented in context with the use of a computerized tool for developing the design.

Having developed the software for computer-aided design, ETS measurement scientists have begun, with some success, the process of determining what to measure during the process of using these tools. NCARB wishes to use this design in a high stakes certification examination but student architects and working architects could benefit from such a design tool in their own office merely as a tool. as the ETS measurement scientists determine ways for scoring excellent, acceptable and mediocre responses to a design specification, this information could be used to provide instructional feedback. The scoring system for excellence in design is referenced to process, not to norms that rank people. Therefore, the processes that receive higher value could be suggested for architects still in training and the system could provide practice and feedback in real-world tasks, including tasks like those to be given later on the certification

[76] Mark M. Sebrechts, L. LeClaire, L.J. Schooler, and Elliot Soloway, Toward generalized intention-based diagnosis: GIDE, In R. C. Ryan (Ed.), Proceedings of the 7th National Educational Computing Conference, Eugene, Oregon, International Council on Computers in Education, 1986, pp. 237-244.

[77] Marc M. Sebrechts, Randy Elliot Bennett, and Donald A. Rock, Machine-scorable constructed-response quantitative items: agreement between expert system and human raters' scores. ETS Research Report RR- -90, Oct. 1, 1990.

examination.

ETS is following the diffusion research model by first substituting a computer for the paper and pencil design tasks now used in a high-stakes exam, and graded holistically by humans. This pioneering work nevertheless paves the way for a new approach to interactive help system that uses unobtrusive process measures during tool use.

The second example involves a simple word processing test with the ability to identify different operations and tactics used by the skilled word processing professional. The test can give feedback that is diagnostic by identifying specific processes that are and are not mastered. Since word processing manuals and tutorials are quite complete for today's products, the test can be used for screening, but beyond high stakes uses it will certainly be used in the field as a learning progress help system. The test consists of standardized reference tasks (assigned papers) to type. Students can take the test as often as they like. Feedback and instruction can be added to the interaction that occurs while students are working on the standardized tasks. The sequence of such standardized typing assignments could take students through all of the features of the word processor, providing an integration test with built-in diagnostic feedback during the process.

These two examples of unobtrusive measures taken during tool use are illustrative of a new form of measurement wholly different from previous measurement approaches. Process measurement can be totally integrated with instruction, as well as with production and performance in both standardized and creative tasks.

Representation Processing Software.

There is a growing family of computer tools that can be used for developing student exhibits. These tools offer significant opportunities for integrating assessment with instruction. It is useful to call these tools representation processing software, because each of them takes some form of information representation (i.e., text, graphics, or numerical representations) and provides processing algorithms and tools to prepare, shape and present representations using each of these forms of communication.

Text and Linguistic Processors. Tools are now available and will be improved in the future. These include outline processors for creating and thinking about what is to be written, word processors and linguistic processors. Linguistics processors are the least familiar, but a few are available in the form of grammar checkers. It is possible to process information at a much deeper level than spell checking or grammar checking, by looking at sentence structure and style.

In a project sponsored by ETS, a computational linguist, Eldon Lytle[78] applied some of his linguistic processing software, "WordMap," to essays that had been graded holistically by teachers. The WordMap scores were as good as a second reader in general, and could therefore be used as one "reading" of an essay. In low-stakes assessment this would save teacher and student time. Fellow students and teachers could perform the definitive readings. In other unpublished research, Lytle has pioneered the development of scales of linguistic maturity. By processing selections of standard writing, a score can be generated that compares the student's style on a continuous scale referenced to products typical of each grade. Selections from great authors are placed on the scale, as well as selected writing examples of individuals known within local schools or communities. The resulting scale has potential as a learning progress growth scale.

Text Search and View Processing Software. Large textual databases now available on magnetic or optical memories allow students to formulate queries that direct the computer to search through texts and retrieve interesting passages from several books about a topic of interest. They learn to assemble key words, perform searches, and to extract and think across many documents. They may use "view processing" software to assemble search material into an organized framework. It may be an outline framework, a hypertext framework, or the process material may be clipped and put into a word processor

[78] Lytle and Breeland, 1989. AERA paper.

for document integration.[79]

Graphics. Programs like MacPaint, MacDraw, PC Paint, SuperPaint, and others are now widely available. These programs give computer users the ability to develop graphic productions of great sophistication. Libraries of "ClipArt" are available and can be used quickly by experienced users to assemble quality presentations. Presentation software packages are available that combine both text and graphics in View Graph formats. These displays can be presented directly from the computer screen or printed out and turned into handouts and transparencies. These programs allow icons and line drawings to be assembled, and going beyond, they allow photographic images to be scanned in with grey scale and even color. More sophisticated tools allow video processing. These production tools bring with them good inherent motivation. Students are able to produce exhibits that incorporate excitement, interest, humor and self-expression. Moreover, students are challenged by the highly professional media they see around them. Graphical information processing systems may be instrumented with unobtrusive measures and with learning progress feedback.

Numbers and Formulas. Schools now use data bases and data sets for experiments in a variety of fields. Tools include spreadsheets for numerical calculations, and software tools that allow mathematical transformations, proofs of systems of equations, etc. For developing mathematical models, special software tools like Stella[80] are used to develop general systems models with underlying mathematical modeling components.

Propositional Knowledge. Systems that allow verbal and logical statements to be processed using theorem checkers and theorem provers are not widely available in schools, but will be in the future. These systems will be the descendants of current expert system shells. Some college classes use these tools currently.

Process measures during tool use offer the best CEA alternative for assessing strategies. Strategies are defined within the context of optional ways to use these tools, and also within sets of tasks having a common goal structure. The use of the tools explicate the strategy and tactics in a way that is accessible to measurement and feedback. Needless to say, a great amount of research and development is needed before promising new kinds of process measures during tool use will enjoy widespread use in schools.

CEA systems utilize computers in the processes of development, distribution and analysis, as well as administration. Technology trends in these areas will be discussed in the following section.

TRENDS IN DEVELOPMENT, DISTRIBUTION, AND ANALYSIS PROCESSES FOR FUTURE CEAs

Using Computers in the Development of Measurement Instruments

Current mark sense testing has a very well elaborated development process with supporting technologies. Future CEA systems will require an infrastructure with new roles and skills for personnel, as well as new technological tools. A selected few are described below:

o Job Analysis. New methods of job and task analysis will emerge. Many new jobs of the future will involve technological tools in some way . Any of these technological tools can be augmented to collect process data from both expert and developing job incumbents. Our prediction is that computerized methods of collecting data from technology-using job incumbents will make it easier to determine the types of tasks engaged in by them, and the different standards for performing these tasks. In

[79] Jostens Learning Corp. now offers Compton's Encyclopedia on-line for students to use in this manner.

[80] Stella user's manual.

addition to online data collection, Job analysts/observers will be able to capture information about technical and non-technical tasks in video, audio and symbolic forms via notebook computers in real time, at the location where work is performed.

o Developing Items and Tasks. Item banks are now in use, but these are relatively primitive. Although some use database technology, they have generally not integrated items fully with multi-media editing capabilities. They have made the text easy to edit, but not the graphics. Item bank development systems are currently designed to produce printed tests, not multimedia interactive tests. At this writing, no test development group has integrated graphics, animations, audio, and video into their production systems. The trend toward multimedia will change the nature of item banks, along with the associated item editors, and types of tasks selected from these banks will improve and become fully integrated across media.

The trend toward performance tasks and exhibits will revolutionize the meaning of the term "item bank". Many of the tasks stored in these banks will not be items at all, but standardized performance situations. Some of these will consist of rather complex simulations with associated scoring protocols and keys. Thus, the new item banks will become sophisticated computer programs with associated multi-media files -- simulation tasks written as objects, or composed of objects using object-oriented programming techniques. There is potential for using these banks in multiple situations so long as the new use can be accompanied by the development of a new scoring and feedback protocol.

o Data Collection for Pretesting, Norm Development and Item Calibration. One of the most significant breakthroughs in CEA could come through the development of a distributed network of temporary data collection sites where response data, (including pre-response data) could be collected rapidly and transmitted electronically overnight to permanent data centers. This concept is depicted in Scenario 2, and its significance for educational assessment cannot be overstated. Achievement test norms now used in schools are rarely updated more often than once every seven years. As the population changes, these norms become less and less appropriate as interpretive frameworks for assessment. With on-line sampling for norm data, it could be done annually.

It is a very costly process to prepare pretest materials for schools willing to function as data collection sites, and to administer these instruments with the necessary quality control. Other costs are transmitting the data back, scoring, and processing. The idea expressed in Scenario 2 enables a sample of schools to provide responses to new sets of items and tasks, providing an exciting opportunity to participate in an important national project, yet also get feedback useful to their own local programs. National assessments like the NAEP survey could take place much more quickly once the infrastructure of test schools (and the means for shifting and adjusting the sample on an a regular basis) is developed. In addition, state and local norms could be developed that more accurately reflect the dynamic and changing demographics of communities.

Decentralized Development.

One of the most remarkable prospects for materials development will be the prospects for decentralizing these complex and difficult development processes. We have already witnessed this rapid decentralization in the evolution of microcomputer-based desktop publishing systems. As "desktop multi-media production systems" become less expensive and more widely available, this phenomenon will further extend these capabilities. The net effect is that schools, academic research facilities, and commercial development and consumer groups can create the interdisciplinary teams necessary for innovative and high quality materials development. Editable curriculum and assessment materials will be developed and adapted to the unique needs of communities.

A vital link in making these local adaptations progressive and evolutionary in their quality will be the widespread implementation of formative evaluation software in schools where computer-aided, interactive teaching and assessment materials are used.

New Possibilities for Formative Evaluation and Improvement. With local file servers and networking to all of the interactive learning/testing stations within a school, response records on all students can be kept for assessment purposes. Some process measures include response times for items and completion times for lessons and modules. Software can summarize this data dynamically so that it is available quickly to help determine what is or is not working.

Ideally, there will also be incentives to encourage schools and districts to send their data back to the developers of the systems so that those developers will be able to obtain sufficient response data to effectively revise the expensive and complex systems for really improved future editions. The data which is actually used should come from carefully sampled and representative locations. Since there will be an increasing demand for localization of at least some of the modules in a system, it will be important for local users to have access to formative evaluation data. Richard Snow and Ellen Mandinach[61] have made this point well in their important document on integrating assessment and instruction.

"Formative evaluation, and adaptation to local circumstances, is the sign quo non of research and development on systems that integrate instruction and assessment (SIIA). We do not imagine SIIA as being 'designed' and then 'implemented,' as those terms are typically used. Rather, we expect that each such system will evolve in its time, place and domain as a function of continuous monitoring and tinkering, even though each may start from the rough common scheme described in this report...

It follows that SIIA design should contain provisions for continuous monitoring and evaluation of its own functioning in each usage."[62]

Trends in Distribution for CEAs.

Distance learning, facilitated by telecommunications technology, represents a promising trend that can influence the distribution of CEAs. As mentioned in Section 2, it has already been used for distance registration and for testing at permanent testing centers. Moving beyond the distribution of assessments to fixed as well as temporary locations, telecommunications will be used to make interactive help systems available for many subject areas to people in schools and in nontraditional learning environments. Help systems do not require data security considerations inherent in High Stakes assessment, thus providing rich opportunities for innovative, interactive multi-media applications.

CEA and help systems can also be integrated into retraining programs for teachers, district administrators, etc. CEA, and help systems applications can benefit from distance learning as a means of in-service training. Broadcasts featuring outstanding teachers could be followed up locally with practice using built-in professional development materials in the help systems.

A combination of magnetic or portable digital media -- like magnetic disks, optical disks or tape, offer additional delivery alternatives to mailing printed tests or to using telecommunications as a means of distributing tests to remote sites. Portable media is also an alternate means of sending response data back to central sites. Portable media in combination with express mail and overnight delivery options may sometimes offer a more cost effective assessment distribution alternative.

Trends In Analysis and Record-keeping.

Among the problems encountered in implementing computerized instruction and assessment systems is the overwhelming amount of data that can be recorded about each student response. Coherent models for recording, prioritizing, organizing, retrieving and analyzing all the new kinds of time and response data have not yet been developed. Advances in measurement science are closely tied to advances in the technologies and analysis methods associated with record-keeping.

[61] Snow, R. E., and Mandinach, E. B., Integrating Assessment and Instruction: A Research and Development Agenda, 1989, Pre-Publication Draft.

[62] Richard E. Snow and Ellen B. Mandinach, Integrating Assessment and Instruction, a Research and Development Agenda. Research report, Educational Testing Service (and press).

As measurement science begins to develop models utilizing time as a key variable, analysis and record-keeping systems will be revised to deal with time data. The time variable has the prospect of becoming one of the most important tools for researchers and evaluators interested in improving CEA systems and associated instruction. The time and the utilization of lessons and smaller modules within them can be obtained from on-line systems. The interpretation of time statistics becomes very important. Some modules will stand out as being extremely time-consuming and hence, probably more difficult than is desirable, and potentially confusing. The combination of time, errors, use, and avoidance of certain features become data for formative evaluators of the future who are researching effective methods and how to improve them. When learner choice is involved, avoidance becomes a potentially useful variable for the assessment of conative objectives. Why were certain carefully crafted and important modules attended to briefly then later avoided?

The trend toward portability can affect analysis and record- keeping as well. In the early 1980s, two of the co-authors proposed the use of a credit-card sized data storage card for all military personnel. Imprinted with registration information, the card would be capable of containing all information gathered about training experiences, assignments, and achievements gained throughout each recruit's career. Several technologies are now available for such data cards. One technology places a microchip into the card which contains memory that can be written, but not erased. The familiar magnetic strip credit card or teller machine card does not store enough information to make this application a reality, but the optical card which uses a photographic emulsion that can be written on with a laser (but not erased, an important part of this application) is now a reality and is entering into a variety of commercial applications. For example, Blue Cross/ Blue Shield uses such cards for the storage of personal medical information among its clients.

CEA SYSTEMS WILL EXERT A TRANSFORMATIONAL INFLUENCE ON ALL PERSONNEL IN THE EDUCATIONAL ENTERPRISE

The widespread introduction of CEA systems will mean that the work roles of the materials developers, the administrators, the teachers, and the students will be changed. CEA systems, especially when integrated intimately with instruction, are transformational in nature. They can become an important tool for the current quest for restructuring; indeed, improved outcome assessment is essential to many calls for substantial educational restructuring.[83]

Much has been said about how students will play a more dynamic and active role as performers in complex tasks, as producers of their own products, as monitors of their own growth using assessment information that shows where they are and how they are progressing. Teachers fulfill a more challenging and responsible role as assessors, managers with more degrees of freedom, and professionals equipped with powerful instrumentation, much as other professionals who have previously passed out of a prominently labor-intensive model.

Among these many role changes, the role of the teacher is the most central and the most significant.

How CEA Systems Will Impact the Role of Teacher

Previous substantial technology interventions into the classroom have fallen short of promised benefits until teachers could accept, then adopt, new roles. Managing a conventional classroom group leads teachers to a set of mental models about what is important and what is possible. Group discipline and group progress through the topics in a district's curriculum guide are all-important. Dispensing information is the way to *cover* this curriculum. Attention to each individual to assess their progress and make sure they are ready to move on is not possible. It has been unthinkable, so has not been thought about. With CMI systems and their successors, Integrated Learning Systems, these priorities change. It is now possible to accomplish things not in the teacher's previous repertoire. These novelties require role changes. It is also necessary to learn new technical skills.

[83] Kerns, Thomas, and Bowser, John books.

Both the shift in role perception and the need to develop new technical skills takes time. The former seems to take from 3 to 5 years with an Integrated Learning System, in the authors' combined experience.

CEA, especially help systems that integrate assessment and instruction using computer technology, provides even greater demand for both role shifts and new technology skills. In this pressure to learn new ways, teachers and other educators will be experiencing the very driving force of technology transforming jobs that demands that they prepare their students for such changes. But many teachers themselves are recent products of a snippet curriculum. Not all have the thinking skills society wants them to inculcate in the young. Not all of them have the confidence or willingness to learn new computer skills, measurement know-how and assessment sophistication, ability to use representation-processing software tools, or bar-code readers, or portable computer scoring systems for holistic scoring. These future technologies move into the inner sanctum of the classroom. At least ILS systems remain in a room down the hall, and have a trained systems administrator on duty to take care of the hardware and software complexities.

So if role changes for ILS systems take 3-5 years, how long will it take for CEA help systems, a more sophisticated member of the ILS family?

No one knows yet, but the prospect is not as bleak as it appears to be. Not all teachers have to adopt systems like those described in this section at once. The early-adoptor teachers are already eager and willing to try it. They have already thought of their own solutions to many of the cognitive, conative, measurement, and management problems CEA systems are designed to solve. America will learn from these early adopters, especially if we back them up with formative evaluation and look for slow, progressive evolutionary improvement, not revolution.

Policy issues are found on both the educational side and the technology side. Educational policies have been discussed. Technology policies and standards in telecommunications, computer use, and software can affect the ability of researchers, developers, and practitioners to use these technologies in the manner proposed.

Barriers to Widespread Implementation of CEA Systems

The largest barrier to the introduction of CEA systems using interactive computers is the implementation of the infrastructure. The hardware must be installed in schools, and the professionals must be trained to adapt to new technological aspects of their jobs. They must be helped to expand and define their own roles in a progression that will take several years, and will find them performing different activities and different roles than at present. Besides the infrastructure barrier, several other barriers currently exist.

Hardware Compatibility Barriers.

There are two major computer hardware standards which exist in schools. These are represented by Apple and IBM microcomputers and their compatibles. Only recently have hardware and software solutions been provided which allowed Apple programs to run on IBM hardware and vice versa for IBM programs to run on Apple hardware. An additional barrier is the limited processing speed, text, and graphics capabilities of many of the current microcomputers implemented in schools. Many of the school computers have eight bit computers, while the state-of-the-art is 16 bit and 32 bit microcomputers. Text resolution on many of the computers in schools is limited to 40 characters per line and 24 lines per page. This limitation is very restrictive for reading comprehension items using extended text passages or for problem solving items. The graphics resolution of many of the microcomputers is limited to 320 x 240 pixels per screen. This graphics limitation is very restrictive for presenting realistic line and shaded graphics. A further barrier is the incompatibility between monochrome and color displays. Most significant for CEA, record keeping requires networked PC's. Stand-alones will have limited use in CEA.

Hardware Availability Barriers.

Although many states and districts are implementing computer technology as rapidly as is financially feasible, there are still a large number of states, districts, and schools which have limited or no availability of computers for use in computerized testing. Industry standards typically recommend that schools have a minimum of thirty-two computer workstations for a school of 500 students and sixty-four computer workstations for a school of 1,000 students. This standard may be too low. The hardware availability barriers are further compounded for disadvantaged schools and districts unless they use the Federal Chapter 1 and Chapter 2 monies to purchase computer hardware for curriculum and computerized testing needs.

Software Availability Barriers.

There are only a small number of companies that currently provide professional computerized testing software and computerized tests for schools. Thus, even if schools have the hardware available, it is unlikely that they will have computerized testing software or computerized tests which they can administer on their hardware. There is a need for the development of professional standards for import and export of computerized testing software, item banks, and computerized tests.

As discussed above, the integrated learning system market is approaching one billion dollars, and has provided a foundation for the evolution to integrated learning and assessment systems. In this environment, it is probable that computerized testing will make significant inroads by offering computerized performance simulations for higher order thinking skills, tool software that assists in the development and evaluation of student products, and helps for teachers and students to integrate instruction and assessment.

Experience and Communication Barriers.

Many of the nation's teachers and students have never taken a computer-administered measurement instrument or used a computerized test development and administration system. Teachers and older students are reluctant to embrace new educational technologies until they have had sufficient personal experience with the technology. In general, the younger the student, the less the fear of trying it.

There is a very small group of educational measurement professionals who are conducting research, developing computerized testing products, and preparing research and dissemination papers on computerized testing. The ERIC Clearinghouse on Tests, Measurement, and Evaluation at the American Institutes for Research collects bibliography references and abstracts of research on computerized testing and computerized adaptive testing. The Buros Institute of Mental Measurements, at the University of Nebraska, has also a computerized testing review and information exchange system. Still, there are very few educational measurement professionals, and even fewer educational professionals (superintendents, principals, curriculum personnel, and teachers) who are acquainted with the research base, computerized testing demonstration systems, and sources for seeking further information.

SUMMARY

In section III, more evidence has been presented to support the policy recommendations made in the last section. In particular, new technologies are making possible much more variety and potential cost-effectiveness in the administration of performance tasks and student exhibits. Group technologies and portable systems will offer even more options. Holistic process measures, judged by teachers and students, are now possible using the intermediate products from representation processing software tools. In the future, direct process measures will be possible during student engagement with computerized performance tasks and during tool use.

Such process measures can be used to integrate assessment with instruction in the unobtrusive and continuous fashion recommended in this paper. These developments in technology present an opportunity to move in the directions of new assessment practices capable of assessing higher level educational objectives. Computer technology provides the new variable in the decision logic of how to achieve the educational goals our country needs.

It is true that extensive R&D, and expensive and long-term professional development of educators is needed, but the current scene already includes both performance tasks and tool use for student projects. The recommendation is not to wait years for the R&D to be completed, but to encourage many projects and put each of them into a Formative Evaluation improvement loop. Progressive and evolutionary transformation of America's educational systems can only come about through slow processes that build from where we are now. Progressive transformation can only be accomplished by the professionals who will use the new formative evaluation and individual assessment tools.

SECTION IV: FINDINGS AND RECOMMENDATIONS

FINDINGS: CURRENT USES OF COMPUTERS IN ASSESSMENT.

The findings are summarized in the Executive Summary, where they are organized around the stages of technology diffusion: substitutive, incremental, and transformational. Another perspective on current uses and trends in CEA is given by the following key findings:

1) Purposes of Assessment:

High-stakes assessment dominates in both individual assessment and program evaluation. Today's testing industry exists primarily to justify high-stakes decisions. Decision-makers want to assure their clientele that their decisions are valid and fair, and are based on a dependable scientific standard. Thus they are willing to pay for testing of individuals (or require individuals to pay for it themselves). They will sometimes pay for summative evaluations, even though program evaluation takes a secondary position to individual assessment.

2) Objectives of Assessment:

Despite the fact that educators universally desire the equivalent of integration, transfer, and creative production objectives, scaffolding objectives dominate assessment practice. Essays and other student projects are emphasized where more than lip service is paid to this desire. Another positive trend is to use more extended problems in math and science, much longer and more challenging than the short standardized items, but class schedules interfere if problem engagement takes more than forty minutes.

The introduction of computerized educational assessment was found to offer fundamental hope for changing these finding. In Section III, it was shown how curriculum developers lack a coherent set of achievement constructs and ways of measuring them. In writing their objectives for a curriculum, developers have depended on verbal objectives that lead to conventional test items. These do not take long to administer, do not require expensive and time-consuming holistic assessment, and are familiar and accepted. The print delivery system does not offer powerful and feasible ways to achieve objectives more complex than scaffolding objectives. Networked computer workstation, however, can implement and disseminate standardized performance tasks, opportunities for student exhibits and portfolio management. They can enhance the practice of discussing and even measuring process during tool use. These measurement methods offer the possibility of getting at the higher objectives of integrated performance, creative production, and strategy improvement. In so doing, they also offer the possibility of achieving the conative objectives of interest, motivation, and persistent effort.

3) Measurement Practices for Individuals:

- o Standardized tests based on short items dominate. However, holistic scoring of student products is gaining momentum, despite high costs. Raters must be trained, and common standards must be set for multiple raters.

- o Integrative performance tasks have taken a strong hold in Military and Industrial training and assessment. CEA systems offer an attractive way to extend these benefits to schools.

- o Tools to aid in producing student products offer new opportunities both to increase this form of learning activity and to assess process through unobtrusive measures taken during tool use. Assessment of strategies is possible with this new measurement practice.

4) Program Evaluation:

Summative evaluation predominates over formative, but it is recommended that this balance shift the other way. Formative Evaluation is typically less favored, perhaps because it is expensive and labor intensive if done as a separate activity, and requires a continuing expenditure of development funds to act on the results and make the improvements indicated. Learning progress assessment is used in some classrooms, but is neither formalized as a process nor computerized to any extent. Most evaluations use individual measures of achievement, summarizing and aggregating them at schoo', districts, state, and national levels. It is recommended that direct measures of system performance be utilized more fully. This is possible with integrated learning/assessment systems and with group-based assessment systems. In addition, long-term measures beyond immediate post-tests should be investigated; the results are often surprising.

5) CEA Administration Alternatives:

- o Scannable answer sheets dominate, with all the back-up technologies these imply. Because of the inherent strengths and probable continuing viability of answer sheet testing, it is vital to extend assessment options by using a wider variety of item types, now possible using bubble answer sheets, but underutilized. Many more item types are possible using higher resolution scanners.
- o Student exhibits are receiving increased emphasis, both as standardized and unstandardized asks. The expertise of students in assessment of their own products and those of others increases when portfolio methods are used.
- o Educational workstations and special workstations in labs are becoming more widely available. When connected to networks, these open up new possibilities for powerful forms of CEA all four of the measurement methods.
- o Portable response units and notebook computers offer the opportunity to put the advantages of educational workstations into the classroom.
- o Computer graphics integrated with video and with active response offers new possibilities for visualizing difficult concepts. Computer graphics based on scientific models has made it possible to visualize for scientists, and in the future for students, phenomena from the very small to the very large.

6) The Role of the Testing Industry in Introducing Computerized Educational Assessment.

It was found that, for a number of reasons, testing companies are not likely to be the leaders in introducing computerized assessment or instruction integrated with assessment. Quite aside from the constraints operating on the testing companies, the installation of hardware in schools is being justified for instruction, with assessment as an afterthought. Thus it is more likely that companies and state organizations concerned with instruction and with strong savvy in the key technologies will provide the leadership. These companies have people who have learned the lessons from Computer-managed and Computer-Assisted Instruction. They have instructional developers on their staffs who know how to develop excellent interactive instruction.

Organizations strong in instructional technology using interactive, networked computers presently are deficient in measurement expertise; however, it is becoming a competitive advantage to have strong measurement and assessment components and to integrate these well with instruction. Therefore, these organizations will find ways around the lack of measurement expertise. People can be hired, but few have the combination of measurement, instructional, and interactive computer know-how. Good people can be developed over time. Testing companies may help provide the measurement expertise if they are willing to work with the instructional technology companies as a new kind of client.

It is the clients of the testing companies who must provide the leadership in Computerized Educational Assessment if it is to come from the testing infrastructure at all. Some of the old clients, in particular state education agencies, membership organizations made up of schools and colleges, and professional societies are well positioned to be leaders in computerized assessment, and even assessment integrated with learning and instruction. Whether or not these existing users of the services of testing companies step forward as leaders or not, a variety of new clients are emerging.

Testing companies and the measurement profession that stands behind them have been leaders in developing and promoting professional standards for test use. The new technology companies, including ILS companies, software companies, and hardware companies are generally not familiar with these standards, and may not be sympathetic with them. It will be a challenge for policy to see that high standards are maintained and that the profit motive will not overwhelm attention to standards. Maintaining high standards is not without its costs.

7) The evolution of the new Infrastructure.

At the end of section II considerable attention was given to the current testing infrastructure and the needed new one. This new infrastructure is distributed to schools and colleges as hardware installed in classrooms and labs. This equipment will be useless unless the human capabilities are developed to use it. It behooves the developers of new educational technologies to attend to the traditions, patterns, and habits of operation of schools and teachers if they want to develop a market that more easily appreciates and understands their products.

The ILS companies have developed their products around the concept of individualized instruction -- learning labs where the students work independently. Individualized instruction is alien in the main to schools as they now operate. Extending ILAS concepts into classrooms through group-paced, presenter-controlled activities offers great promise for the future. The technologies of interactive computers for display on projectors or large monitors, and the technologies of portable individual response units seem naturally suited for the introduction of ILAS concepts into classrooms. The goal of better learning and instruction through integrated, unobtrusive, and helpful assessment into classrooms is worth the effort it will take.

THE ARGUMENT OF THIS PAPER IN A NUTSHELL

A shift has occurred in the purposes of testing. Selection and judging are not the central problems for educational measurement any more. Improving achievement and progress for a demographically diverse student population, to standards achieved by only a few now, has become the central problem.

Measurement relates to this need to enhance learning progress in a fundamental way. Achievement constructs are needed, and associated standards of excellence that go beyond the simplistic minimum competencies that are current in so-called reform efforts.

Measurement and assessment are fundamental to all occupations and professions because decision making is fundamental to all, and decision-making depends on sensitive judgments based on appropriate and accurate information.

Improvements in educational measurement practice continue to be vital because it is essential to know what to measure and how to measure each important construct. A generally acceptable set of achievement constructs with associated measurement methods effective for each type of achievement needs to be developed. Table 8 is a simple model for this concept. Good measurement practice also requires a way to manage the measurements and other data. Mainframe Computers have been used in this task in the past. Decentralized computers in classrooms and labs will be used in the future. Good measurement practice requires in addition a way to keep the measures up to date. Systems that perform all of these functions are presently called Integrated Learning Systems (ILS), but they are weak on measurement, so-so on management, and getting stronger on instruction. They provide the framework for introducing and using good measures, and when they do, they could more appropriately be called Integrated Learning and Assessment Systems (ILAS).

Good assessment relates to the nation's need to enhance learning progress in a fundamental way. The high-level objectives that go beyond scaffolding knowledge currently require sensitive interpretation of student performances and products, and of the processes involved in each. Those who perform the assessments are the teachers and the more advanced students. No outside agency, human or machine, can do it. These people require a clear conception of the standards of excellence. Teachers need to keep improving their ability to relate student performances and products to these standards. They also need to keep improving their ability to provide instruction, hints, and helps that will enable all students with the desire to learn to achieve. Assessment also involves wise decision-making, including decisions about how to guide students into paths that will promote learning progress. In short, good assessment practice involves both knowing how to interpret data in context with other information and knowing how to use the interpretations in day-to-day educational decision making.

In contrast to measurement as practiced in some other occupations, educational measurement plays a peculiar and skewed role. There is a lack of consensus on what should be measured, a lack of measurement instrumentation (except not fully appropriate printed instruments), and a lack of an infrastructure for obtaining the appropriate measures and keeping them up-to-date. Educational measurements are skewed toward measurement for High-Stakes decisions imposed on the teachers and/or the learners from outside.

Computer systems can integrate administration of measurement instruments, presentation of instructional materials, record-keeping, and management of records and of instructional activities. Computers interact, so they can provide optional advice and help from moment to moment, not just at formal workshop inservice sessions. This can provide a new kind of growth environment for classrooms and learning labs. The growth environment is important for teachers as professional instructors and assessors, and for students as future contributors and assessors. The introduction of appropriately configured computer systems is the most promising way now visible for providing this new environment for productive learning, teaching, and growth for all people in the system.

Top-down implementation models that do not involve the teachers, other educators, and students in a process of slow growth will probably not work. Even if the perfect Integrated Learning and Assessment System were fully developed today, it would be rejected by users unless they were able to develop their own roles. Therefore, it is better to start where the users are now, and introduce formative evaluation methods, enlisting the aid of the users to improve the systems over subsequent generations based on formative evaluation data. Formative evaluation then, is more than a better method of program evaluation; it is a fundamental aspect of implementation strategy.

RECOMMENDATIONS

Recommendations Dealing With the Purposes and Methods for Testing in the Schools

Recommendation One: Greatly increase the frequency and variety of help services compared to high-stakes assessments, but balance the two.

1.1 Through research and development funding and policy support, encourage the development of high-stakes measures that can be integrated and correlated intimately with the curricula at a few key milestones and that measure more integrated and comprehensive achievement objectives than the scaffolding objectives now common.

1.2 Reduce the use of item tests and other scaffolding level tests for incremental grading practices and shift the burden, both of grading students and holding teachers accountable to these more carefully developed, more integrated, and less narrowly construed high-stakes measures.

1.3 Support research, development, and implementation of help systems that integrate measurement with instruction.

1.4 Provide training for both teachers and students in holistic assessment of standardized performance tasks and student exhibits and of the intermediate products and processes leading up to these final products.

Recommendation Two: Increase the frequency of formative evaluation, and provide funding and incentives to use the evaluation data for on-going improvement of educational programs.

2.1 Encourage research and development on direct measures of system utilization and performance, as well as measures of student achievement, also on methods for summarizing these measures, with attention to student privacy issues, for the use of developers in revising and improving curricula and assessment materials.

2.2 Make summative and formative evaluation continuing processes rather than special projects, while utilizing the carefully developed high-stakes measures proposed in Recommendation 1.1 for continuing summative decision-making.

2.3 Since summative decisions are of the go/no-go variety, and do not contribute to improvement, administrators should be encouraged to provide resources to improve based on formative evaluation.

With the shrinking teacher force projected over the next decade, increased use of technology may be the only cost-effective way to implement new learning-oriented approaches to education. But new systems must be installed and improved over time as the teaching force learns new roles and new technical skills.

Recommendation Three: Increase the Use of Alternate Methods of Assessment That Require Human Judgment

3.1 Support research, development, and implementation which places more emphasis on measurement methods of performance tasks, exhibits, and process measures during tool use, and thus more teaching and learning of the higher-order constructs of integration, creative production and strategies.

3.2 Increase the variety of item types for the item tests that measure scaffolding knowledge. This kind of knowledge can conveniently be measured by verbal multiple choice or short-answer items, but there are many new item types that can be delivered on both paper and pencil and computer that should be further developed and used.

3.3 Encourage the development and use of standardized performance tasks, such as simulations of complex and realistic situations, games, and tool-like laboratory environments that allow students to make choices and decisions and observe the realistic consequences. Utilize these computerized performance tasks to assess integrated performance objectives.

3.4 Support research and development leading to automated scoring schemes for computerized performance tasks, but in the meantime encourage the use of holistic scoring techniques on the part of both students and teachers.

3.5 Support and encourage the development of methods for using student exhibits as a part of assessment integrated with instruction. These methods include portfolio management procedures and require that students and teachers be taught holistic scoring methods for essays, performances, student-generated experiments, etc. Use these measurement methods as part of assessment of creative production objectives.

3.6 Encourage the development of measurement methods that assess <u>intermediate products</u> students develop by holistic methods. This recommendation applies both to performance tasks and exhibits, and emphasizes process measures. Use these methods to assess strategy improvement objectives.

3.7 Support the development of automated computer scoring of process measures during student use of computer tools like word processors, spreadsheets, or presentation packages. Investigate recording student responses while they are using computer tools, and develop software to provide hints and helps on strategy at the moment of need.

Recommendation Four: Foster new item types and uses of portable answer media in order to utilize the current testing infrastructure more creatively.

4.1 Provide and encourage R&D funding and seek to install policy-based incentives for organizations that perform testing to encourage them to introduce new item types beyond multiple choice. New item types should offer new methods for obtaining student responses on answer sheets, using high resolution scanners and other technologies.

4.2 Encourage the development of assessments integrated with instruction that utilize the new answer sheet item types and use them for practice and feedback in connection both with group presentations and with individual seatwork.

Recommendation Five: Encourage the development of the new, localized infrastructure of Integrated Learning and Assessment Systems, and the coordinated evolution of central sites for development and for R&D.

5.1 Encourage the further development and implementation of computer-based approaches to integrating learning and assessment. Encourage many approaches rather than one kind of concept and product.

5.2 Encourage the evolution of what might be called an ILAS industry through funding R&D centers, regional labs, non-profit organizations, state agencies, and other organizations to develop integrated instructional and assessment materials, and to conduct research and development.

5.3 Support research, development, and implementation for group-oriented systems that integrate instruction and assessment. Needed technology includes projectors and software for teachers to present excellent instructional materials with integrated group-paced assessments. The display capabilities should include color, graphics, video, and audio. The student response entry technology that needs to be developed includes response pads, infra-red linkages, and student-oriented portable computers.

Recommendation Six: Encourage the professional development of teachers and other professionals who are knowledgeable and skilled about both the human judgment and the echnical aspects of CEA, and are skilled at integrating assessment with instruction.

6.1 State agencies, school districts, and professional associations should encourage conferences, publications, and program development activity to effect this recommendation.

6.2 Provide incentives for colleges of education to introduce new programs for the development of professionals who can provide skilled holistic assessment and who can integrate assessment with instruction.

6.3 Support research and development which will lead to "built-in" computerized consultants and advisors into integrated learning and assessment systems to provide a continuous professional growth program for teachers who are users of the systems. The consultation and advice will occur at the moment of need during the school day, not limited to summers, released time or weekend workshops

Policy Recommendations:

Recommendation Seven: Federal and state policy should both provide R&D funds and stimulate private sector investment in improving technology-based assessment practices.

Effective policy that focuses on stimulating R&D and innovative product development is recommended to update enhanced high stakes assessment options, also to stimulate the creation of the high quality help systems proposed in this paper. These types of measurement systems can no longer be left up to the teachers to develop in their spare time. The types of research and development are detailed in Recommendations One through Six.

Recommendation Eight: Encourage the maintenance and improvement of high professional testing standards for CEA systems.

Computerized testing requires faithful adherence to the established professional standards for test construction and evaluation, standards for test use, and standards for administrative procedure. These standards have been codified in the following professional references: Standards for Educational and Psychological Testing, Guidelines for Computer-Based Tests and Interpretations, Code of Fair Testing Practices in Education, and Technical Guidelines for Assessing Computerized Adaptive Tests.

Professional assessment standards require careful focus on technical issues of test validity, test reliability and errors of measurement, scaling, norming and equating of computerized tests as well as issues related to computerized test administration, scoring and reporting, and protecting the rights of test takers. The purpose of the standards is to provide detailed criteria for the evaluation of tests, testing practices and effects of test use. Standards help to ensure that test developers and administrators focus on issues of validity, comparability, equity, ethical issues, bias, and confidentiality of scores.

The advent of Learning Progress help systems may require modification of these standards. For example, reliability is vital in a high-stakes admissions test, and must be bought at a price -- more items and more testing time. Admissions is a high-stakes decision, but whether to take a learning module or not is a low-stakes decision that can easily be corrected if a decision didn't work. It is not worth the time and cost of the high-stakes standard for reliability. By contrast, construct validity -- *what is really being measured and learned* -- is of utmost importance in both kinds of systems.

8.1 Policy should maintain a continuing emphasis on Equity in evaluating CEA systems. The intent is to provide equality of educational opportunity for disadvantaged groups as well as advantaged groups. If the more advantaged districts and schools purchase and implement computerized testing technology, while the disadvantaged districts and schools do not, then there will be an inherent inequality of educational opportunity. When economically advantaged students have greater access to computerized technology in their homes than economically disadvantaged children, then there will be some level of inequality of educational opportunity.

8.2 Policy should emphasize fairness issues in the development and implementation of CEA systems. The code of fair testing practices in education was developed to safeguard the rights of test takers. With computerized tests, the intent of the fair testing practices is to provide a fair and appropriate test for each examinee whether the test is administered by computer or by paper and pencil. This translates into support for studies of item or test performance differences for a particular kind of test for members of age, ethnic, cultural or gender groups in the population of test takers. Such research should be designed to detect and eliminate aspects of test design, content, or format that might bias test scores for particular groups.

8.3 Support studies that assure statistically that two score scales are equivalent (equating studies) in order to establish the degree of comparability of scores from computerized tests and paper-administered tests when both forms are administered to the same population.

IN CONCLUSION

America's needs are great, but American ingenuity has been at work over-time to come up with technological tools and ideas for using them in education. The tools and ideas cut across many disciplines, but there are other ideas, embodied in systems that integrate management, instruction, and assessment into coherent and usable systems. Systems exist now for equipping learning and testing rooms as permanent centers. Systems also exist and new ones are predicted, that reach out into the classrooms and provide support for presentation, assessment, records management, and practice.

There is much work to be done in the areas of science, technology, infrastructure building, and support as educators' roles evolve. America has met challenges before, and can and will meet this one.